

---

**Supplementary information**

---

# Global chemical effects of the microbiome include new bile-acid conjugations

---

In the format provided by the authors and unedited

Robert A. Quinn, Alexey V. Melnik, Alison Vrbanc, Ting Fu, Kathryn A. Patras, Mitchell P. Christy, Zsolt Bodai, Pedro Belda-Ferre, Anupriya Tripathi, Lawton K. Chung, Michael Downes, Ryan D. Welch, Melissa Quinn, Greg Humphrey, Morgan Panitchpakdi, Kelly C. Weldon, Alexander Aksenov, Ricardo da Silva, Julian Avila-Pacheco, Clary Clish, Sena Bae, Himel Mallick, Eric A. Franzosa, Jason Lloyd-Price, Robert Bussell, Taren Thron, Andrew T. Nelson, Mingxun Wang, Eric Leszczynski, Fernando Vargas, Julia M. Gauglitz, Michael J. Meehan, Emily Gentry, Timothy D. Arthur, Alexis C. Komor, Orit Poulsen, Brigid S. Boland, John T. Chang, William J. Sandborn, Meerana Lim, Neha Garg, Julie C. Lumeng, Ramnik J. Xavier, Barbara I. Kazmierczak, Ruchi Jain, Marie Egan, Kyung E. Rhee, David Ferguson, Manuela Raffatellu, Hera Vlamakis, Gabriel G. Haddad, Dionicio Siegel, Curtis Huttenhower, Sarkis K. Mazmanian, Ronald M. Evans, Victor Nizet, Rob Knight & Pieter C. Dorrestein<sup>✉</sup>

## Supplemental Information

### Global Chemical Impact of the Microbiome Includes Novel Bile Acid Conjugations

Robert A. Quinn<sup>1,2</sup>, Alexey V. Melnik<sup>1</sup>, Alison Vrbancac<sup>3</sup>, Ting Fu<sup>4</sup>, Kathryn A. Patras<sup>3</sup>, Mitchell Christy<sup>1</sup>, Zsolt Bodai<sup>5</sup>, Pedro Belda-Ferre<sup>3</sup>, Anupriya Tripathi<sup>1,3</sup>, Lawton K. Chung<sup>3</sup>, Michael Downes<sup>4</sup>, Ryan D. Welch<sup>4</sup>, Melissa Quinn<sup>6</sup>, Greg Humphrey<sup>3</sup>, Morgan Panitchpakdi<sup>1</sup>, Kelly Weldon<sup>1</sup>, Alexander Aksenov<sup>1</sup>, Ricardo da Silva<sup>1</sup>, Julian Avila-Pacheco<sup>7</sup>, Clary Clish<sup>7</sup>, Sena Bae<sup>8,9</sup>, Himel Mallick<sup>7,8</sup>, Eric A. Franzosa<sup>7,9</sup>, Jason Lloyd-Price<sup>7,9</sup>, Robert Bussell<sup>10</sup>, Taren Thron<sup>11</sup>, Andrew T. Nelson<sup>1</sup>, Mingxun Wang<sup>1</sup>, Eric Leszczynski<sup>6</sup>, Fernando Vargas<sup>1</sup>, Julia M. Gauglitz<sup>1</sup>, Michael J. Meehan<sup>1</sup>, Emily Gentry<sup>1</sup>, Timothy D. Arthur<sup>3,7</sup>, Alexis C. Komor<sup>3</sup>, Orit Poulsen<sup>3</sup>, Brigid S. Boland<sup>12</sup>, John T. Chang<sup>12</sup>, William J. Sandborn<sup>12</sup>, Meerana Lim<sup>3</sup>, Neha Garg<sup>13,14</sup>, Julie C. Lumeng<sup>15</sup>, Ramnik J. Xavier<sup>7</sup>, Barbara I. Kazmierczak<sup>16</sup>, Ruchi Jain<sup>16</sup>, Marie Egan<sup>17</sup>, Kyung E. Rhee<sup>3</sup>, David Ferguson<sup>6</sup>, Manuela Raffatellu<sup>3</sup>, Hera Vlamakis<sup>7</sup>, Gabriel G. Haddad<sup>3</sup>, Dionicio Siegel<sup>1</sup>, Curtis Huttenhower<sup>7,8</sup>, Sarkis K. Mazmanian<sup>11</sup>, Ronald M. Evans<sup>4,21</sup>, Victor Nizet<sup>1,3,19</sup>, Rob Knight<sup>3,18,19, 20</sup> and Pieter C. Dorrestein<sup>1,3,19</sup>

<sup>1</sup>Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA

<sup>2</sup>Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI

<sup>3</sup>Department of Pediatrics, University of California San Diego, La Jolla, CA

<sup>4</sup>Gene Expression Laboratory, Salk Institute for Biological Studies, La Jolla, CA

<sup>5</sup>Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA

<sup>6</sup>Department of Kinesiology, Michigan State University, East Lansing, MI

<sup>7</sup>Broad Institute of MIT and Harvard, Cambridge MA 02142

<sup>8</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115

<sup>9</sup>Department of Immunology and Infectious Diseases, Harvard T.H. Chan School, Boston, MA

<sup>10</sup>Department of Radiology, University of California San Diego, La Jolla, CA

<sup>11</sup>Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA

<sup>12</sup>Division of Gastroenterology, Department of Medicine, University of California San Diego, La Jolla, CA

<sup>13</sup>School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA

<sup>14</sup>Emory-Children's Cystic Fibrosis Center, Atlanta, GA

<sup>15</sup>Department of Pediatrics, University of Michigan, Ann Arbor, MI

<sup>16</sup>Department of Internal Medicine, Yale School of Medicine, New Haven, CT

<sup>17</sup>Department of Pediatrics, Yale School of Medicine, New Haven, CT

<sup>18</sup>Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA

<sup>19</sup>UCSD Center for Microbiome Innovation, University of California, San Diego.

<sup>20</sup>Department of Engineering, University of California, San Diego

<sup>21</sup>Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037

## Table of Contents

1. Methods
2. Supplementary Data
3. Supplementary Tables
4. Supplementary NMR spectra
5. Supplementary 3D mouse model and x,y,z coordinates.

## Methods

**Animals.** Germ-free (GF) C57Bl/6J mice were generated via caesarian section and microbiologically-sterile animals were cross-fostered by GF Swiss-Webster dams at the California Institute of Technology. GF animals were housed in open-top caging within flexible film isolators (Class Biologically Clean; Madison, WI) and maintained microbiologically sterile, confirmed via 16S rRNA PCR from fecal-derived DNA and culture of fecal pellets on Brucella blood agar or tryptic soy blood agar (Teknova; Hollister CA) under anaerobic and aerobic conditions, respectively. The same mice as the GF were grown under non-GF conditions. Conventionally-colonized specific pathogen free (SPF) mice (C57Bl/6J) were housed in autoclaved, ventilated, microisolator caging. All animals received autoclaved food (LabDiet Laboratory Autoclavable Diet 5010; St Louis, MO) and water *ad libitum*, were maintained on the same 12-hour light-dark cycle and housed in the same room of the facility. All animal husbandry and experiments for this component were approved by the California Institute of Technology's Institutional Animal Care and Use Committee (IACUC). All animal dissections and sample collection for the GF and SPF mouse aspect of the study were carried out at University of California at San Diego under IACUC approval, protocol S00227M. For MRI imaging, a female, C57Bl/6 mouse, 8 weeks of age, was obtained from Jackson Laboratory and housed with food and water *ad libitum*. For metabolome and microbiome studies, four germ-free (GF) and four specific-pathogen-free (SPF) female 8-week-old C57Bl/6J mice were acquired from the California Institute of Technology's vivarium. Samples of the food the animals were provided were also collected and analyzed (GF were fed LabDiet 5010 and SPF were fed LabDiet 5053, LabDiet, St. Louis, MO).

An additional 24 male ApoE knockout mice in the C57BL/6J background raised for use in a study of hypoxia on the murine microbiome according to the methods of Tripathi et al. 2018<sup>1</sup> were also analyzed in this study for the effects of high-fat-diet and feeding <sup>13</sup>C-Phe on the new bile acids. The fecal samples collected, and the data presented here were not published in that study and approved under IACUC S05534. The source data from this murine experiment is available online.

**Human Sample Collection:** Fecal samples were collected from two separate pediatric cystic fibrosis patient cohorts for detection of novel bile acids. One sample set was collected from patients at the Rady's Children's Hospital in San Diego, CA using dual fecal swabs according to the procedure outlined in the American Gut Project<sup>2</sup> under IRB approval #160034. The second collection was done on CF patients with pancreatic sufficiency, without pancreatic sufficiency and healthy controls at Yale New Haven Hospital (New Haven, CT) under IRB approval #1206010476 according to the procedure outlined in<sup>3</sup>. Two separate IBD cohorts were also analyzed for the presence of the novel bile acids. The first for detection through GNPS data searching according to the American Gut Project fecal collection protocols and the second for searching a completely different patient cohort with different

collection methods and mass spectrometry analysis from the human microbiome project 2 (HMP2) according to the methods of<sup>4</sup>. The UCSD stool sample collections from patients with IBD were collected as part of the UCSD IBD Biobank under IRB #131487. Human infant fecal samples were collected at the University of Michigan under IRB #103575.

**3D Model Generation:** A female, C57Bl/6J mouse, 8 weeks of age, was euthanized using carbon dioxide inhalation and then immediately brought to the UCSD Center for Functional MRI. The MRI images were acquired on a Bruker 7T/20 MRI scanner using a quadrature birdcage transceiver. A 3D FLASH protocol with TE/TR=6 ms/15 ms and matrix size 128x64x156 was used, prescribing a field of view to match the body size. The dicom files from the mouse MRI were imported into the Invesalius software<sup>5</sup>. In Invesalius, the dicom files were visualized as stacked images through the axial, sagittal and coronal slices. Organs of interest were then traced in each slice according to their best visualization in the different viewpoints. The tracing was done using 'create new mask' feature in Invesalius using the manual edition mode. The brush feature was used to trace the outline of each organ of interest in the appropriate slice, stack by stack, until the entire organ was outlined through all slices in each orientation such that its outline was smoothed and did not bleed into other organs. Numerous iterations of this process led to the mapping of each organ through the MRI stacked images. The 'Configure 3D surface' feature was then used to translate the 2D stack tracings into a 3D image of each organ. This was completed for all organs sampled except for blood, fecal and skin samples, successively, until an entire 3D-model of all organs of interest to this study was built. Blender (<https://www.blender.org/>) was used to smooth the model and color each organ differently, enabling better visualization of the different organs and organ systems. Blood and skin samples were not mapped onto the model and a representative fecal sample was added after MRI modeling using Invesalius to allow mapping to a theoretical fecal sample.

**Sample Collection.** Mice were euthanized via carbon dioxide asphyxiation. Prior to dissection, external sites including the skin (left and right flank), ears, mouth and feet were sampled using a cotton swab with vigorous contact for 5 seconds. Blood was collected via cardiac puncture using a 22-gauge needle and 1 ml syringe. Mice were then sterilely dissected under open flame using straight scissors and fine forceps that were cleaned with 70% ethanol (v/v) between handling of each organ. The following organs were dissected: Adrenal gland, bladder, brain, cecum, cervix, colon, duodenum, esophagus, foot, gall bladder, heart, ileum, jejunum, kidney, liver, lung, ovaries, spleen, stomach, thymus, trachea, uterus and vagina. Additional samples were collected using swabs including skin, ear, foot, and mouth. The sample collection order is shown in table S1. Sections of each organ were made using sterile razor blades, with the number of sections listed in table S1. The

123 liver and lung were sectioned into their corresponding lobes (Liver: right and left median lobes, right  
124 and left lobes and caudate lobe; Lung: superior lobe, middle lobe, inferior lobe, post-caval lobe and  
125 left lung lobe). The heart was sectioned into left and right ventricle and left and right atrium. Each  
126 kidney was sub-sectioned by targeting the outer cortex and inner medulla. The uterus was  
127 subsampled by collecting each left and right uterine horn and oviduct and a single sample of the  
128 uterine fundus. The brain was subsampled by collecting the left and right cerebellum and cerebrum.  
129 The GI samples were sectioned into 6 equal length pieces based on the full length of each GI section  
130 (including 6 sections of the cecum). Margins of the duodenum and jejunum were determined at the  
131 site of the suspensory muscle of the duodenum. The junction of the jejunum and ileum was estimated  
132 as 6 cm proximal to the cecum based on previously reported lengths<sup>6</sup>. The GI samples were not  
133 cleaned or flushed prior to sample collection. The spleen (4 sections), pancreas (3 sections), adrenal  
134 gland (2 sections), and vagina (2 sections) were also sectioned into equal length pieces according to  
135 size (Table S1). It took approximately 45 minutes to fully dissect each mouse immediately after  
136 euthanasia. Four stool samples were also collected from each group of mice from the bedding of the  
137 sterile shipping containers immediately after arrival in the UCSD analysis laboratory. With such  
138 collection method it is not known which mouse produced which stool sample. Food samples fed to  
139 both GF and SPF mice were also collected and analyzed. Sample collection for the additional  
140 published murine studies were completed according to<sup>1,7</sup>. In addition, fecal samples were collected  
141 from mice fed a high-fat diet starting at 10 weeks and compared to animals fed the control normal  
142 chow diet according to the methods of<sup>1</sup>. The data from<sup>1</sup> was not published as part of that manuscript.

143  
144 **Sample Processing:** All samples were contained in 2 ml sterile Eppendorf® Biopur® Safe-  
145 Lock tubes, wet tissue mass recorded, and then frozen at -80°C until metabolite and DNA extraction.  
146 For the swab samples, the wooden end of the swab was cut off with scissors, added to a  
147 microcentrifuge tube and 1 ml of PBS was added. After thawing, all of the non-swab samples were  
148 diluted in a 1:10 mass:volume in sterile phosphate buffered saline. A Qiagen (Qiagen Inc., Valencia,  
149 CA) 5 mm stainless steel bead was added to each tube and the samples were homogenized in a  
150 Qiagen TissueLyzer II homogenizer at a frequency of 20/s for 5 min. After homogenization two  
151 aliquots of 50 µl of the homogenate or PBS/swab mix was added to separate 96-well deep well plates,  
152 one for metabolite extraction and one for DNA extraction. Metabolites were extracted from the  
153 samples in the 96-well deep well plate by adding 200 µl of LC-MS grade 70% methanol in LC-MS  
154 grade water and vortexing each plate for 5 seconds. Samples were left to extract overnight at 4°C and  
155 then spun down to pellet debris in a 96-well plate Sorvall® Legend centrifuge at 2500 rpm for 1  
156 minute. DNA was extracted from the homogenized tissue according to protocols benchmarked for the

157 Earth Microbiome Project (EMP) found here: [http://www.earthmicrobiome.org/emp-standard-](http://www.earthmicrobiome.org/emp-standard-protocols/)  
158 [protocols/](http://www.earthmicrobiome.org/emp-standard-protocols/)<sup>8,9</sup>.

159

160 **LC-MS/MS Mass Spectrometry:** A 50 µl aliquot of the extracted sample in methanol was  
161 added to a 96-well plate and diluted with 150 µl of LC-MS grade methanol containing 2 µl of ampicillin  
162 MS internal standard. The chromatographic separation was conducted on a ThermoScientific  
163 UltraMate 3000 Dionex UPLC system (Fisher Scientific, Waltham, MA USA) with eluent subsequently  
164 electrospray ionized and analyzed with a Bruker Daltonics® MaXis qTOF mass spectrometer (Bruker,  
165 Billerica, MA USA). Metabolites were separated using a Kinetex 2.6 µm C18 (30 x 2.10 mm) UPLC  
166 column containing a guard column. Mobile phases A 98:2 and B 2:98 ratio of water and acetonitrile,  
167 respectively, containing 0.1% formic acid and a linear gradient from 0 to 100% for a total run time of  
168 840 s at a flow rate of 0.5 mL min<sup>-1</sup> were used. The mass spectrometer was calibrated daily using  
169 Tuning Mix ES-TOF (Agilent Technologies) at a 3 mL min<sup>-1</sup> flow rate. A lock mass internal calibration  
170 was used by soaking a wick with hexakis (1H,1H,3H- tetrafluoropropoxy) phosphazene ions  
171 (Synquest Laboratories, *m/z* 922.0098) located within the source. Full scan MS spectra (*m/z* 50 –  
172 2000) were acquired in the qTOF and the top ten most intense ions in a particular scan were  
173 fragmented using collision induced dissociation at 35 eV for +1 ions and 25 eV for +2 ions in the  
174 collision cell. A data dependent automatic exclusion protocol was used such that an ion was  
175 fragmented upon its first detection, then fragmented twice more, but not again unless its intensity was  
176 2.5x the previous fragmentation. The isolation width was dependent on *m/z* with a 4 *m/z* isolation for  
177 50 *m/z* to 8 *m/z* at 1000 or higher. This exclusion method was cyclical, being restarted after every 30  
178 seconds.

179 Mass spectrometry data for the mice fed a high-fat diet compared to normal chow for 10  
180 weeks was generated separately from this study on a ThermoScientific™ qExactive™ mass  
181 spectrometer according to the procedure of<sup>1</sup>. The mass spectrometry data generation for the HMP2  
182 (PRISM and iHMP datasets) was completed also on a ThermoScientific™ qExactive™, but in negative  
183 mode as described in<sup>10</sup>. These methods are less likely to capture known microbiome derived volatiles  
184 such as short chain fatty acids.

185

186 **Metabolomics Data Processing and Analysis.** Each LC-MS/MS file in the Bruker format (.d)  
187 was converted to .mzXML format using the Bruker® DataAnalysis 'Process with Method' batch script.  
188 Lock mass calibration was applied during conversion to aid in mass accuracy. The .mzXML files were  
189 uploaded to the UCSD MassIVE data storage server for GNPS analysis. The entire dataset is publicly  
190 available and found under the ID MSV000079949. In addition, the area under curve feature  
191 abundances were calculated in batch for all files using the Optimus<sup>11</sup> software based on the OpenMS

192 feature finding algorithms<sup>12</sup>. The Optimus parameters were as follows: *m/z* tolerance 15.0 ppm, noise  
193 threshold of 3000, retention time tolerance of 20 s, intensity factor compared to blanks at 3.0, and a  
194 feature observation rate of 0.01. The data was then trimmed to contain information only from 60 s to  
195 550 s of the run during the linear gradient; this removed wash steps programmed into the run at the  
196 start and end of the chromatographic program. The feature abundances were normalized to the total  
197 ion current (TIC) in each sample for statistical analysis by dividing the area-under-curve abundance  
198 for each feature in each sample by the total ion current of that sample (TIC-normalization). For organ-  
199 by-organ beta-diversity analysis the features present in individual organs were extracted as separate  
200 feature tables and any features not present at all in a particular organ were removed. Additional data  
201 for the HFD study<sup>1</sup> was generated with a ThermoScientific™ qExactive™ mass spectrometer, and  
202 processed using the mzMine software<sup>13</sup> with the feature table TIC-normalized. Parameters were as  
203 follows: MS<sup>1</sup> minimum threshold of 10000 counts, MS<sup>2</sup> threshold of 5000 counts, a mass tolerance of  
204 0.03Da and retention time tolerance of 0.2 min. The data was deconvoluted, deisotoped and filtered  
205 for compounds present in at least 3 samples. This additional metabolomics dataset is publicly  
206 available under MassIVE ID MSV000082480.

207 Molecular networking was performed on GNPS with the GF and SPF mice samples separated  
208 from each other and from blank and quality control samples using the group-mapping feature. The  
209 molecular networking and MS-cluster parameters were as follows: parent and fragment ion mass  
210 tolerance 0.05 Da, minimum cosine score of 0.7, minimum matched fragment ions of 4, and a  
211 minimum cluster size of 4 (to minimize detection of more rare nodes found in few samples). The  
212 library search parameters of the molecular networking search were a minimum-matched peaks of 4  
213 and a cosine score of 0.65. Any library hits from the results were inspected directly between the  
214 spectrum and query and are considered level two according to the metabolomics standards  
215 consortium guidelines<sup>14</sup>. The estimated false discovery rate (FDR) for spectral matching is 4.1% under  
216 our search parameters<sup>15</sup>. A link to the full data molecular network used for statistical analysis and  
217 annotation is available  
218 here <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9ea760fb819449d7bc7aca8fec07bd8d>.

219 Meta-mass shift chemical profiling of chemical transformations between nodes was done using  
220 the method of<sup>16</sup>. Briefly, all nodes unique to either GF or SPF were searched for an edge connection  
221 to a node from one or the other groups (GF to SPF, SPF to GF, GF to shared or SPF to shared). This  
222 represented a molecule unique in either GF or SPF mice that was related to a molecule in the other  
223 group, indicating it was modified in sterile or colonized mice. In each instance, the mass gain or loss  
224 relative to the unique node was recorded along with the spectral count for each node as a measure of  
225 its abundance. Mass differences were binned into known molecular modifications within a 0.03 Da  
226 window as described in<sup>16</sup> with the addition of unique modifications relevant to this dataset, such as

227 saccharides. All other unknown mass shifts were ignored. Mass shifts that were counted included H<sub>2</sub>  
228 (m/z2.02) acetyl (m/z42.05), methyl (m/z14.02), H<sub>2</sub>O (m/z18.01), C<sub>2</sub>H<sub>4</sub> (m/z28.03) O (m/z16.00),  
229 CH<sub>2</sub>O (m/z30.91), NH<sub>3</sub> (m/z17.03), C<sub>2</sub>H<sub>2</sub> (m/z26.02), C (m/z12.01), C<sub>2</sub> (m/z24.02), CH<sub>4</sub> (m/z16.04),  
230 SO<sub>3</sub> (m/z79.96), C<sub>4</sub>H<sub>8</sub> (m/z56.06), 2H<sub>2</sub> (m/z4.03), C<sub>2</sub>H<sub>6</sub> (m/z30.05), CH<sub>2</sub>O<sub>2</sub> (m/z46.01), CO<sub>2</sub>  
231 (m/z43.99), OH (m/z17.01) and sugars corresponding to C<sub>6</sub>H<sub>10</sub>O<sub>4</sub> (m/z146.06), C<sub>6</sub>H<sub>10</sub>O<sub>5</sub> (m/z162.05),  
232 C<sub>5</sub>H<sub>8</sub>O<sub>4</sub> (m/z132.04) and 2 glycone units C<sub>12</sub>H<sub>18</sub>O<sub>11</sub> (m/z338.09). The spectral counts for node  
233 representing the specific modification were summed and plotted as total spectral counts for that  
234 modification in GF and SPF mice as either mass gains or losses.

235

236 **16S rRNA Gene Amplicon Sequencing of Mouse Samples:** On all murine samples  
237 collected both GF and SPF and control samples of solutions and swabs underwent DNA extraction,  
238 16S rRNA gene variable region 4 (V4) PCR and amplicon preparation for sequencing according to  
239 protocols benchmarked for the Earth Microbiome Project (EMP) found  
240 here: <http://www.earthmicrobiome.org/emp-standard-protocols/><sup>9,17</sup>. The microbiome data was  
241 processed through the Qiita software (qiita.ucsd.edu). The data was demultiplexed, reads trimmed to  
242 150 bp, and Deblur<sup>18</sup> was used to de-noise the data into sub-OTUs (sOTUs). The resultant .biom files  
243 were used for downstream analysis with QIIME<sup>8</sup>. To create a phylogenetic tree for UniFrac<sup>19</sup> analysis,  
244 deblurred sOTU sequences were inserted into the annotated Greengenes<sup>20</sup> tree with SEPP<sup>21</sup> and  
245 taxonomy assigned using the corresponding taxonomic label on the internal node where the  
246 sequence inserted. The microbiome data is available at (<https://qiita.ucsd.edu/>, study ID:10801).

247

248 **3D Mapping in *'ili*:** Metabolomics and microbiome data were mapped onto the 3-D mouse  
249 model by recording the location of the sampling and orientation of each sample in the model  
250 according to the methods described in<sup>11</sup>. Some organs only contained one sample (bladder, blood,  
251 cervix, gall bladder and thymus) all other organs contained 2-6 samples and the actual location of the  
252 dissected sample was mapped to the appropriate point representing that same sample in the 3D  
253 model. The point mapping was done using the GeoMagic® Wrap software. The full .stl model of the  
254 laboratory mouse was loaded into GeoMagic Wrap and the location of each sampling point was  
255 selected with the 'points' tool (available as supplemental data). The x,y,z coordinate information in the  
256 model from all points was then exported as a .csv file for matching to its representative sample in the  
257 metabolomics or microbiome data (available as supplemental data). Sub models of different organ  
258 systems were also created in the same manner to aid visualization, such as the GI tract and liver.  
259 Mapping to these models was done as described for the full model. For *'ili* visualization, the matching  
260 samples for the 4 GF and 4 SPF mice were averaged and a new feature or OTU table created based  
261 on these mean abundances. This feature table was then matched to the x,y,z coordinates from the



262 model according to the correct sample. This OTU or metabolite feature table was then uploaded into  
263 the *'ili'* software simultaneously with the mouse model. This enabled automatic mapping of the  
264 abundance of a microbial or metabolite variable to the point representing its collection location in the  
265 GF and SPF mouse 3D-model. Visualization in *'ili'* was done using a linear scale with the *'viridis'* color  
266 map and automatic min/max mapping was selected.

267

268 **Statistical Analysis Of the Mouse Data:** The microbiome .biom table and metabolome  
269 feature table were analyzed using principal coordinate analysis after calculation of a distance matrix  
270 between all samples. Alpha diversity of the metabolome data was calculated using the Shannon-  
271 Weiner index on the TIC-normalized feature table from the murine GI tract in the R statistical software.  
272 The microbiome distance matrix was generated using the unweighted UniFrac distance<sup>22</sup> in QIIME  
273 and QIIME2. Beta-diversity of the microbiome data was calculated on a feature table rarified to 500  
274 reads per sample to enable visualization of GF and sterile samples which had a low number of 16S  
275 rDNA gene reads. Repetition at higher read thresholds produced very similar results for the SPF  
276 samples, as expected from prior studies. The metabolomic beta-diversity was calculated using the  
277 Bray-Curtis dissimilarity. The resulting distance matrix was visualized using principal coordinates  
278 analysis (PCoA) and each sample highlighted by either GF/SPF or organ source for both groups of  
279 mice. To assess the overall similarities between the metabolome of murine organs the Bray-Curtis  
280 dissimilarity was calculated between all paired samples (compared for the same subsection location  
281 for the same organ) between the GF and SPF states for all mice and these dissimilarities were  
282 averaged per organ and plotted with notch plots. This same comparison was done within GF and SPF  
283 groups to determine the level of variation for mice of the same classification. In addition, the within  
284 group variation was compared between GF and SPF mice separately in the same manner.

285 To determine the number of unique metabolites between GF and SPF in each organ molecular  
286 networks were built with the same above parameters for samples from each of the 29 organs. The  
287 molecular networking data was then downloaded from GNPS and the source of each node as GF or  
288 SPF was tabulated. A spectrum was considered unique to either class of mice only if it was detected  
289 in at least 3 out of 4 individual mice sampled per category. Each instance of these unique nodes was  
290 counted and reported as a percentage of the total number of nodes from each organ and as the total  
291 number of nodes per organ to visualize abundance. This was also done at the level of each individual  
292 mouse comparison to obtain a degree of variation in the overall unique metabolite differences.

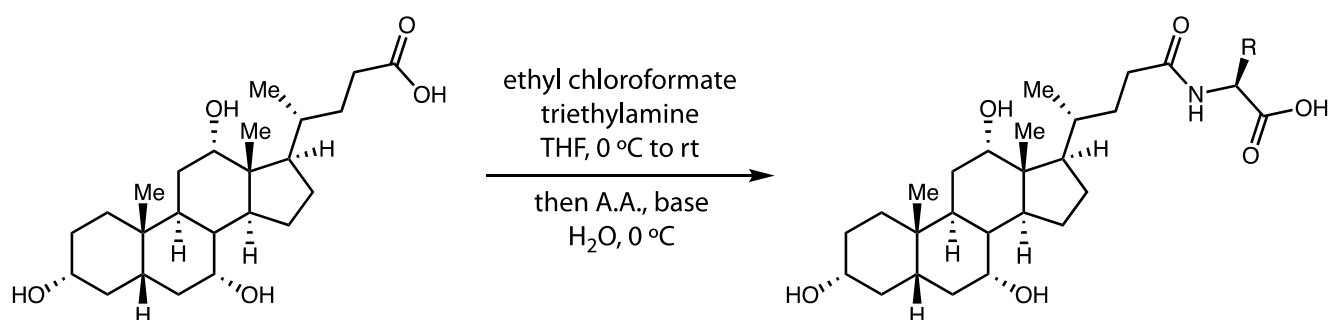
293 To visualize the effect of the GF or SPF classification on the gut metabolomic data a random  
294 forests classification was run on all GI tract samples (including the esophagus) and the variable  
295 importance for classification of each metabolite was determined. The random forest analysis was  
296 done using 5000 trees with the R-statistical package *'random forests'*. The variable importance plot

was then computed for the metabolites most differentiating the GF and SPF states of the animals. These variables of importance were analyzed for known compounds in GNPS and their molecular family memberships. The 30 most differentially abundant metabolites according to their variable importance were then visualized using a stacked bar graph showing their relative abundance to the entire metabolome. This enabled visualization of the changes in the most differential metabolites through the GI tract and an indication of how abundant these differential metabolites were overall. The Shannon-Weiner index of diversity was calculated on the entire metabolome from each GI tract associated sample using the R statistical software. The mean Shannon-Weiner diversity for each sample location was visualized for the two groups of mice through the GI tract. The Mann-Whitney U-test was used to determine a statistically significant difference ( $p < 0.05$ ) between the Shannon diversity of each GI tract sample collected at the same location between the GF and SPF mice. The microbiome diversity was calculated using the Faith's phylogenetic diversity index in the Qiita software and mean diversity between the four individual mice was presented only for the SPF mice.

Tests of the differential abundance of the novel bile acids between mice fed antibiotics or high fat were done using the Mann-Whitney U-test with a significance level of  $p < 0.05$ . Correlations between the feature abundance of the novel bile acids and bacterial OTUs from the HFD experiment were calculated using the Pearson's correlation on deblurred reads. Reads with the highest correlations were assigned by BLAST to the NCBI nucleotide database with only cultured representatives included in the search.

The alpha diversity of the batch culture experiment was calculated using the Shannon-Index on the deblurred OTU table produced through Qiita and the duplicate sequenced samples were averaged.

**Synthesis of Novel Conjugated Bile Acids.** The procedure was adapted from a previous method by Ezawa et al.<sup>23</sup> Cholic acid (100mg, 0.25mmol, 1 eq.) was dissolved in THF (4.9mL, 0.05mM) and cooled to 0°C in an ice water bath with stirring. Ethyl chloroformate (28μL, 1.2 eq.) was added followed by triethylamine (41μL, 1.2eq) and the reaction stirred for 0.5 hours cold. After complete conversion of the starting material by TLC a cold, aqueous solution (4.9mL) of amino acid (0.37mmol, 1.5 eq.) and base (0.37mmol, 1.5 eq.) is added in one portion. The reaction is stirred for 1 hour at 0°C to completion. THF is removed under reduced pressure and 2M HCl is added to acidify to  $\text{pH} < 2$  and a white precipitate appears. The mixture was extracted with ethyl acetate (3 x 20mL), the combined organic layers washed with brine (1 x 50mL), dried over sodium sulfate, and concentrated. Purification was done by column chromatography with 6% --> 18% Methanol/DCM + 1% acetic acid to give the desired product as a white solid.



<u>R (Side Chain)</u>	<u>Base</u>	<u>Isolated Yield</u>
Leucine	NaHCO <sub>3</sub>	79mg (62%)
Isoleucine	NaHCO <sub>3</sub>	74mg (58%)
Phenylalanine	NaHCO <sub>3</sub>	85 mg (63%)
Tyrosine	NaOH	80 mg (57%)
<sup>13</sup> C-Tyrosine	NaOH	137 mg (94%)

331

332

Leucine Conjugate: 62% Yield. Product made using the general procedure. White solid. <sup>1</sup>H

333

**NMR** (600 MHz, MeOD) δ 4.37 (s, 1H), 3.96 (s, 1H), 3.80 (d, *J* = 2.6 Hz, 1H), 3.40 – 3.34 (m, 1H),

334

2.36 – 2.22 (m, 3H), 2.21 – 2.13 (m, 1H), 2.03 – 1.94 (m, 3H), 1.93 – 1.78 (m, 4H), 1.78 – 1.51 (m,

335

10H), 1.47 – 1.27 (m, 5H), 1.15 – 1.06 (m, 1H), 1.04 (d, *J* = 6.5 Hz, 3H), 1.02 – 0.94 (m, 4H), 0.94 –

336

0.89 (m, 6H), 0.71 (s, 3H). <sup>13</sup>C **NMR** (151 MHz, MeOD) δ 176.80, 74.05, 72.87, 69.04, 48.12, 47.49,

337

43.18, 42.99, 41.96, 41.00, 40.44, 36.91, 36.48, 35.90, 35.85, 34.02, 33.33, 31.16, 29.56, 28.73,

338

27.86, 26.13, 24.24, 23.56, 23.16, 21.81, 17.73, 13.00. **M.P.** = 175-178C. **IR** – 3390.24, 2933.2,

339

2868.59, 2426.01, 1634.38, 1464.67. **HRMS** (ESI) exact mass calculated for [M+H]<sup>+</sup> (C<sub>30</sub>H<sub>52</sub>NO<sub>6</sub>)

340

requires *m/z* 522.3789, found 522.3793 with a difference of 0.77 ppm.

341

Isoleucine Conjugate: 58% Yield. Product made using the general procedure. White solid. <sup>1</sup>H

342

**NMR** (599 MHz, MeOD) δ 4.32 – 4.27 (m, 1H), 3.96 (s, 1H), 3.80 (d, *J* = 2.8 Hz, 1H), 3.40 – 3.34 (m,

343

1H), 2.38 – 2.15 (m, 4H), 2.03 – 1.93 (m, 3H), 1.93 – 1.78 (m, 4H), 1.78 – 1.50 (m, 10H), 1.45 – 1.27

344

(m, 4H), 1.26 – 1.19 (m, 1H), 1.11 (qd, *J* = 11.8, 5.6 Hz, 1H), 1.05 – 1.02 (m, *J* = 7.0, 2.0 Hz, 3H),

345

1.01 – 0.90 (m, 10H), 0.71 (s, 3H). <sup>13</sup>C **NMR** (151 MHz, MeOD) δ 176.86, 74.06, 72.87, 69.04, 48.12,

346

47.48, 47.48, 43.18, 42.98, 41.00, 40.44, 38.33, 36.93, 36.48, 35.89, 35.84, 33.87, 33.35, 31.16,

347

29.56, 28.72, 27.86, 26.24, 24.23, 23.17, 17.73, 16.15, 13.00, 11.85. **M.P.** = 144-148C. **IR** – 3392.17,

348

2933.2, 2871.49, 2483.87, 1639.20, 1461.78. **HRMS** (ESI) exact mass calculated for [M+H]<sup>+</sup>

349

(C<sub>30</sub>H<sub>52</sub>NO<sub>6</sub>) requires *m/z* 522.3789, found 522.3792 with a difference of 0.57 ppm.

350

Phenylalanine Conjugate: 63% Yield. Product made using the general procedure. White

351

solid. <sup>1</sup>H **NMR** (599 MHz, MeOD) δ 7.28 – 7.17 (m, 5H), 4.60 (dd, *J* = 8.9, 4.8 Hz, 1H), 3.93 (t, *J* = 2.7

352

Hz, 1H), 3.80 (d, *J* = 2.8 Hz, 1H), 3.40 – 3.35 (m, 1H), 3.22 (dd, *J* = 13.9, 4.8 Hz, 1H), 2.94 (dd, *J* =

353

13.9, 9.1 Hz, 1H), 2.33 – 2.18 (m, 3H), 2.11 – 2.04 (m, 1H), 2.01 – 1.94 (m, 3H), 1.86 – 1.78 (m, 3H),

354

1.76 – 1.63 (m, 3H), 1.62 – 1.50 (m, 5H), 1.47 – 1.33 (m, 3H), 1.21 (m, 2H), 1.09 (qd, *J* = 11.9, 5.3 Hz,

355 1H), 1.02 – 0.95 (m, 4H), 0.92 (s, 3H), 0.68 (s, 3H). <sup>13</sup>C NMR (151 MHz, MeOD) δ 17138.76, 130.28,  
356 129.38, 127.68, 74.04, 72.87, 69.04, 48.02, 47.44, 43.18, 42.97, 40.99, 40.44, 38.47, 36.84, 36.48,  
357 35.89, 35.84, 33.87, 33.23, 31.16, 29.56, 28.66, 27.86, 24.22, 23.16, 17.66, 13.00. **M.P.** = 142-146C.  
358 **IR** – 3395.07, 2934.16, 2865.70, 2494.47, 1638.23, 1455.99. **HRMS** (ESI) exact mass calculated for  
359 [M+H]<sup>+</sup> (C<sub>33</sub>H<sub>50</sub>NO<sub>6</sub>) requires *m/z* 556.3633, found 556.3637 with a difference of 0.72 ppm.

360 Tyrosine Conjugate: 57% Yield. Product made using the general procedure. White solid. <sup>1</sup>H  
361 NMR (599 MHz, MeOD) δ 7.03 (d, *J* = 8.5 Hz, 2H), 6.68 (d, *J* = 8.5 Hz, 2H), 4.52 (dd, *J* = 8.6, 4.8 Hz,  
362 1H), 3.94 (t, *J* = 2.7 Hz, 1H), 3.80 (d, *J* = 2.8 Hz, 1H), 3.40 – 3.34 (m, 1H), 3.11 (dd, *J* = 14.0, 4.8 Hz,  
363 1H), 2.84 (dd, *J* = 13.9, 8.8 Hz, 1H), 2.33 – 2.20 (m, 3H), 2.07 (m, 1H), 2.02 – 1.93 (m, 3H), 1.88 –  
364 1.78 (m, 3H), 1.77 – 1.63 (m, 3H), 1.62 – 1.51 (m, 5H), 1.45 – 1.34 (m, 3H), 1.27 – 1.18 (m, 2H), 1.10  
365 (qd, *J* = 11.8, 5.4 Hz, 1H), 1.02 – 0.95 (m, 4H), 0.92 (s, 3H), 0.69 (s, 3H). <sup>13</sup>C NMR (151 MHz, MeOD)  
366 δ 176.65, 157.21, 131.26, 129.41, 116.10, 74.05, 72.87, 69.05, 48.05, 47.44, 43.17, 42.97, 40.99,  
367 40.44, 37.73, 36.85, 36.47, 35.89, 35.83, 33.95, 33.26, 31.16, 29.55, 28.67, 27.86, 24.23, 23.16,  
368 17.67, 13.00. **M.P.** = 174-178C. **IR** – 3398.92, 2936.09, 2867.63, 1614.13, 1446.35. **HRMS** (ESI)  
369 exact mass calculated for [M+H]<sup>+</sup> (C<sub>33</sub>H<sub>50</sub>NO<sub>7</sub>) requires *m/z* 572.3582, found 572.3584 with a  
370 difference of 0.35 ppm.

371 <sup>13</sup>C<sub>9</sub>, <sup>15</sup>N-labelled Tyrosine Conjugate: 94% yield. Product made using the general procedure  
372 with slight modifications. The reaction time for initial activation of the carboxylic acid at 0°C was  
373 extended from 0.5 h to 2 h. Additionally, following addition of the labelled tyrosine and NaOH, the  
374 reaction time was extended to 2h. The product was obtained as a white solid. <sup>1</sup>H NMR (599 MHz,  
375 MeOD) δ 7.21 – 6.86 (m, 2H), 6.85 – 6.52 (m, 2H), 4.56 (d, *J* = 141.7 Hz, 1H), 3.94 (t, *J* = 3.0 Hz, 1H),  
376 3.80 (q, *J* = 3.1 Hz, 1H), 3.41 – 3.35 (m, 1H), 3.27 – 2.97 (m, 1H), 2.97 – 2.67 (m, 1H), 2.33 – 2.19 (m,  
377 3H), 2.14 – 2.04 (m, 1H), 2.03 – 1.90 (m, 3H), 1.89 – 1.77 (m, 3H), 1.77 – 1.62 (m, 3H), 1.62 – 1.48  
378 (m, 5H), 1.47 – 1.32 (m, 3H), 1.25 – 1.16 (m, 2H), 1.09 (qd, *J* = 11.9, 5.6 Hz, 1H), 1.03 – 0.94 (m, 4H),  
379 0.91 (s, 3H), 0.68 (s, 3H). <sup>13</sup>C NMR (151 MHz, MeOD) δ 157.13, 157.06, 131.21 (t, *J* = 55.2 Hz),  
380 128.95, 116.16 (t, *J* = 62.4 Hz), 74.05, 72.82, 69.08, 49.43, 49.28, 49.14, 49.00, 48.86, 48.72, 48.57,  
381 48.00, 47.39, 43.06, 42.90, 40.88, 40.35, 37.56 (dd, *J* = 47.0, 27.7 Hz), 36.77, 36.41, 35.83, 35.75,  
382 33.84, 33.20, 31.08, 29.46, 28.60, 27.78, 24.19, 23.13, 17.66, 12.98.

383

384 **Novel Bile Conjugates Validation Experiments.** To validate the synthetic standards of the  
385 tyrosine, phenylalanine, leucine and isoleucine cholic and muricholic acids conjugates, the  
386 compounds were dissolved in methanol, diluted to 5 μM and run on the LC-MS/MS method described  
387 above. The data is publicly available under MassIVE ID: MSV000082467. Retention times and  
388 MS/MS spectra were analyzed to verify the molecular characteristics. To determine the approximate  
389 concentration of Phe-chol in the murine GI tract an ileal sample from a GF mouse was spiked with

standard curve of concentrations of pure Phe-chol (non-murine form). Final concentrations of 100  $\mu$ M, 25  $\mu$ M, 5  $\mu$ M, 1  $\mu$ M, 0.1  $\mu$ M and 0.02  $\mu$ M, were directly added to the extracted ileal sample and analyzed with mass spectrometry using the same methods as described above. A standard curve of these concentrations was calculated by plotting the known concentrations to their corresponding area-under-curve (AUC) abundance of the Phe-chol peak. The same AUC abundance was then captured for each sample positive for the molecule in the colonized mice. The concentration in the murine samples was then calculated based on the concentrations of the standard curve. Because isoleucine and leucine cannot be distinguished with MS/MS data, we analyzed the synthetic isoleucocholic acid standard and leucocholic acid standard on an extended gradient HPLC column. The two standards were injected with the jejunum3 sample from mouse SPF2 and subjected to a 40% LC gradient of the same solvents described above with ramp to 40% solvent B at 3 minutes followed by 22 min of ramping to 100% B and then wash steps. The MS/MS method was identical to that described above and retention time differences were recorded between the two chemical standards and the murine sample. To determine whether the base bile acid was either cholic or muricholic acids, the muricholic forms were synthesized according to the supplementary methods in place of cholic acids and all 3 amino acid conjugates of each bile acid backbone were analyzed using the original LC-MS/MS with sample SPF2 jejunum 3, which contained the same molecules detected in the murine gut. Retention time analysis was used to identify whether each molecule in the mouse sample was either muricholic or cholic acid forms. Links to mirror plots showing matches between the novel conjugated bile acids in the murine data and standards are found as follows Leu-chol:

[https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=7ec1a92395c540d78faa34613a64deac&view=view\\_all\\_annotations\\_DB#%7B%22main.Compound\\_Name\\_input%22%3A%22leuco%22%7D](https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=7ec1a92395c540d78faa34613a64deac&view=view_all_annotations_DB#%7B%22main.Compound_Name_input%22%3A%22leuco%22%7D)

Phe-chol

[https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=7ec1a92395c540d78faa34613a64deac&view=view\\_all\\_annotations\\_DB#%7B%22main.Compound\\_Name\\_input%22%3A%22phenylalano%22%7D](https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=7ec1a92395c540d78faa34613a64deac&view=view_all_annotations_DB#%7B%22main.Compound_Name_input%22%3A%22phenylalano%22%7D)

Tyrososocholic acid

[https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=7ec1a92395c540d78faa34613a64deac&view=view\\_all\\_annotations\\_DB#%7B%22main.Compound\\_Name\\_input%22%3A%22tyroso%22%7D](https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=7ec1a92395c540d78faa34613a64deac&view=view_all_annotations_DB#%7B%22main.Compound_Name_input%22%3A%22tyroso%22%7D)

**Mining Public Data Mining on GNPS.** The single spectrum search feature in GNPS (MASST, [https://gnps.ucsd.edu/ProteoSAFe/index.jsp?params=%7B%22workflow%22:%22SEARCH\\_SINGLE\\_SPECTRUM%22,%22library\\_on\\_server%22:%22d.speclibs;%22%7D](https://gnps.ucsd.edu/ProteoSAFe/index.jsp?params=%7B%22workflow%22:%22SEARCH_SINGLE_SPECTRUM%22,%22library_on_server%22:%22d.speclibs;%22%7D)) that allows one to search public MS/MS data through spectral alignment<sup>11</sup> was used to search for the unique amino acid conjugated bile acids in publicly available data. The parameters of the search were as follows: 0.03 Da window of parent mass and fragment ion matching, 0.7 cosine score and a minimum matched peaks of 4 ions. In datasets with a positive hit, the source organism and % of samples positive for

each compound was recorded. Two datasets comprised of LC-MS/MS data analyzed on a Bruker Maxis qTOF from fecal swabs of CF patients (massive IDs MSV000079134 and MSV000082406) were further analyzed according to the metadata of the studies as pancreatic sufficient, insufficient or samples from healthy individuals. The presence of an MS/MS spectrum for each of these classes was tabulated by individual and reported as the percent of subjects positive for each molecule in each class. The results from the MASST searches are available at the following links: [Phe](#), [Tyr](#), and [Leu](#) and can be cloned to search against all public data sets that have become available through GNPS since the these jobs were performed in Sept 2018.

**Development of UPLC-Triple Quadrupole Mass Spectrometry Method for Bile Acids Quantification and Assessment of Matrix Effects.** The above chromatography method used in the murine tissues analysis was transferred to a Thermo Ultimate 3000 UHPLC coupled with a Thermo TSQ Quantum Access Max ESI triple quadrupole (QQQ) system. An identical column, mobile phases, sample injection volume, and column thermostat temperature setting were used as described in the LC-MS/MS section above. However, In order to increase sample throughput, the gradient was slightly modified: gradient elution was set to start with one-minute hold at 5% organic composition, then linearly increase to 90% over four minutes followed by 90% organic content hold for 2 minutes and decrease to 5% and hold for 5 minutes to equilibrate the system before the subsequent injection. The flow rate was set to 0.25 ml/min to match optimal operating regime for the QQQ mass analyzer. The ESI sprayer parameters are summarized in table S5. Multiple reaction monitoring (MRM) transitions were selected to achieve the highest sensitivity and specificity of the targeted molecules. The optimal MRM transitions were selected independently for both regular and stable  $^{13}\text{C}$ -Phenylalanine isotopic labeled synthetic conjugate and  $^{13}\text{C}_9$ ,  $^{15}\text{N}$ -Tyrosine isotopic labeled synthetic conjugate. The Retention Time (RT) and two transitions per molecule were used for the specificity to achieve level 1 annotation<sup>14</sup>. These MRM parameters of all quantified molecules are summarized in Table S5.

**Assessment of Matrix Effects and Measuring of Limit of Detection (LOD) in Different Matrices** Matrix effects on the novel conjugated bile acids from the murine GI tract samples were evaluated to characterize the interferences observed during the untargeted analysis. For this, sample aliquots for each tissue and sample type of GF mice were pooled together, injected, and quantified using an external standard calibration. The calibration curve was created using standards in the 5 ng/ml to 250 ng/ml range. The same samples were also spiked with the 50 ng/ml of each bile acid conjugate and analyzed in identical fashion. Matrix effect values were calculated by comparing the expected value (50 ng/ml) to the difference observed between the assayed samples and the samples with added standard (table S5). As the matrix could affect the LODs due to ion suppression or ion

enhancement; the GF samples (which do not contain the target compounds) were spiked with different concentrations and injected to the HPLC-MS system. Limit of detection was calculated as three times of the standard error of the fitted regression line divided by the slope for each conjugate separately and for each tissue type.

**Quantification of Novel Bile Conjugates in SPF mice with Internal Standard Calibration and Matrix Matched Calibration.** The original samples from SPF mice were re-analyzed with the HPLC-ESI-QQQ targeted quantification method described above with two separate quantification approaches. 1) Internal Standard Calibration: all samples were injected with 2  $\mu$ L of  $^{13}\text{C}$ -Phenylalanine isotopic labeled synthetic bile conjugate and  $^{13}\text{C}_9$ ,  $^{15}\text{N}$ -Tyrosine isotopic labeled synthetic bile conjugate as internal standard mixture (250 ng/ml); mixed in the HPLC injector loop. As the Phe-chol internal standard only had one  $^{13}\text{C}$  modification the natural distribution contribution of the M+1 isotope was corrected during the calculation. 2) Matrix-matched calibration: calibration curves were built to cover the range of 2.5 ng/ml to 1  $\mu$ g/ml for each tissue type by adding external standards into pooled GF mice samples lacking targeted bile conjugates. For both calibrations, linear fitting was used to determine slope and intercept of the calibration curve. These parameters were used to calculate the concentration of unknown samples. The obtained concentrations were then expressed in  $\mu\text{M/g}$  quantities based on masses of original samples.

**Quantification of The Phenylalanine Bile Acid Conjugate Production by Bacterial Strains.** Correlations between the novel bile acids were assessed using the Pearson correlation and mmvec<sup>25</sup>. Cultures of *C. bolteae* CC43 001B and *C. bolteae* WAL-14578 strains were extracted as previously described for the mouse sample processing method. The bile acids in the extracts were quantified using targeted quantification method described above. Elution gradient was set to start with one-minute hold at 5% organic composition, then linearly increase to 90% over four minutes followed by 90% organic content hold for 2 minutes and decrease to 5% and hold for 5 minutes to equilibrate the system before the subsequent injection. The flow rate was set to 0.25 ml/min throughout. The calibration curves were calculated from a range of 0.25 ng/ml to 100 ng/ml with standards.

**Fecal Culture Bioreactor Inoculation.** A 4g stool sample was resuspended in 40mL modified yeast casitone fatty acids media (mYCFA, DMSZ recipe) with 0.25% Antifoam B Silicon Emulsion (Baker) in a vinyl anaerobic chamber (Coy). The resuspension was centrifuged at 500 x g for 5 minutes to pellet solids. The supernatant was decanted through a sterile 70  $\mu\text{M}$  filter. The filtrate was centrifuged at 4450 x g for 10 minutes to pellet cells. The supernatant was discarded, and the pellet was resuspended in 40mL mYCFA. The resuspension was drawn into a 60 mL syringe and injected

496 into a 500 mL vessel of an Infors Multifors 2 bioreactor. The chemostat process parameters was  
497 modified from a previous process developed in<sup>26</sup>. The chemostat volume parameters were; 400 mL  
498 culture volume, 24-hour retention rate, 50 mL/min nitrogen, stirrer at 250 rpm, and 37°C temperature.  
499 10mM stocks of cholic acid, chenodeoxycholic acid, glycocholic acid, Leu-chol, Phe-chol and Tyr-chol  
500 were prepared in 100 µL methanol. 15 µL stocks were added to 12 mL mYCFA. After 11 days of  
501 continuous culturing, 24 mL bioreactor culture was withdrawn and transferred to the anaerobic  
502 chamber. 3mL culture was added to the 12mL mYCFA aliquots with the bile acids, for a total volume  
503 of 15mL and final concentration of 10 µM bile acid. The cultures were vortexed and split into three  
504 5mL aliquots. At time 0 (blanks for each bile acid), 1, 3, 6, 12 and 24 hours, 0.1mL aliquots were  
505 removed from the samples for metabolomics and 16S rRNA gene sequencing.

506 A separate experiment in 96 deep-well plate format was completed in similar fashion with  
507 media formulated according to<sup>26</sup> (designed to mimic human gut contents). A fresh fecal swab  
508 (sampled according to methods from the American Gut Project<sup>2</sup>) was first resuspended in 1x PBS and  
509 then 20 µL of fecal resuspension was inoculated into 500 µL of media in each well. Conjugated bile  
510 acids (Phe-chol, Tyr-Chol, Leu-Chol and Gly-chol) were added to the cultures prior to incubation in  
511 triplicate. The cultures were incubated at 37°C for 48 hours. Both culture experiments (batch culture  
512 and 96-well plate format) were extracted with 70% methanol according to the same methods  
513 described above and analyzed with LC-MS/MS using the same instrument and methods as described  
514 above for GF and SPF mouse studies. The batch culture experiment had microbiome sequencing  
515 completed and analyzed.

516

517 **16S rRNA Gene Amplicon Sequencing of Batch Cultures.** DNA was extracted from the  
518 bioreactor samples using QIAGEN AllPrep 96 PowerFecal DNA/RNA, (QIAGEN custom product #  
519 1114341) with bead-beating on a TissueLyser II (QIAGEN). 16S rRNA gene libraries targeting the V4  
520 region of the 16S rRNA gene were prepared by first using qPCR to normalize template concentrations  
521 and determine optimal cycle number. To ensure minimal over-amplification, each sample was  
522 normalized to the lowest concentration sample, amplifying with this sample optimal cycle number for  
523 the library construction PCR. Four 25 µL reactions were prepared per sample with 0.5 units of  
524 Phusion with 1X High Fidelity buffer, 200 µM of each dNTP, 0.3 µM of 515F (5'-  
525 AATGATACGGCGACCAACGAGATCTACACTATGGTAATTGTGTGCCAGCMGCCGCGGTAA-3')  
526 and unique reverse barcode primer from the Golay primer set<sup>9</sup>. After amplification, replicates were  
527 pooled and cleaned via Agencourt AMPure XP-PCR purification system. Prior to final pooling, purified  
528 libraries were diluted 1:100 and quantified again via qPCR (Two 25 µL reactions, 2x iQ SYBR  
529 SUPERMix (Bio-Rad, REF: 1708880 with Read 1 (5'-  
530 TATGGTAATTGTGTGYCAGCMGCCGCGGTAA-3'), Read 2 (5'-  
531 AGTCAGTCAGCCGGACTACNVGGGTWTCTAAT-3')). Pools were quantified by Qubit (Life



Technologies, Inc.). Final pools were sequenced on an Illumina MiSeq 300 using custom index 5'-ATTAGAWACCCBDGTAGTCCGGCTGACTGACT-3' and custom Read 1 and Read 2 primers mentioned above.

**Farnesoid X Receptor Stimulation from Bile Acids.** Human kidney cell line HeK-293 was obtained from American Type Culture Collection (ATCC CRL-1573, tested for *Mycoplasma* contamination every 6 months). These cells were chosen due to their high transfectability and low FXR expression which allows for a robust signal to noise ratio. These 293 cells were cultured in Dulbecco's modified Eagle's medium/F-12 (DMEM) supplemented with 10% (V/V) heat-inactivated fetal calf serum (FBS) and 100 units/ml penicillin G and 100 µg/ml streptomycin. 10,000 cells were seeded per well in 96-well plates one day before transfection of plasmids. DNA was transiently transfected by Lipofectamine 2000 and Opti-MDM in fasting state. The ratio of plasmid used in per well were 50ng of FXR response element (FXRE)/luciferase reporter plasmid, 10 ng of pCMV-3flag-FXR (human) plasmid, 10 ng of pCMV-RXR (human) plasmid, and 5ng of Renilla luciferase reporter plasmid as internal standard for transfection efficiency. After 12 hrs of transfection, 293 cells were treated with the indicated concentration of bile acids (Phe-Chol, Tyr-Chol, Leu-Chol, CDCA, DCA and T-βMCA.) with FXR synthetic agonist GW4064 as control. Cells were harvested 24 hrs later and lysed with passive lysis buffer (Promega). Luciferase activities were measured by the Dual-Luciferase Reporter (DLR™) Assay kit and read by Luminometer (Perkin Elmer). The final Luciferase activities were normalized by dividing the relative light units by Renilla luciferase activity. Statistical analyses were performed using Prism software. Each dosage was done in 12 replicates.

**<sup>13</sup>C-Phenylalanine Feeding of Mice and Analysis of Fecal Samples.** ApoE<sup>-/-</sup> (Jackson Labs Stock No. 002052) females approximately 16 weeks old were used for this experiment. Fecal pellets were collected from each mouse at baseline (mice were fed regular chow (RC) prior to experiment) and each day after for the duration of the experiment HFD feeding (between 9-11 am each day). Each mouse was housed in an individual cage lined with nestlets. The diet was then shifted to HFD containing 1.25% cholesterol and 21% milk fat (TD96121; Envigo, Madison, WI) at day 0. The overall experiment duration was 9 days with the final stool collection being on day 10. On days 1-3, each mouse was fed the HFD alone. On days 4-6, the experimental mouse was shifted to HFD supplemented with the <sup>13</sup>C-labeled phenylalanine (Catalog # 490091 Sigma-Aldrich) and the control mouse to HFD supplemented with unlabeled phenylalanine. Both groups of mice were shifted back to the HFD without supplemental phenylalanine on days 7-9. The food was prepared as follows each day: each day the HFD pellets were mixed with water from the mouse bottles at 1.5mL water per 10 grams of food to make a uniform slush inside a small dish that is placed on the cage bottom. For days

567 4-6, the amino acid powder at 10 µg/mg was spread on top of the food, water was added and mixed.  
568 Fecal samples were collected from these animals and screened for the production of labeled and  
569 unlabeled Phe-chol. Fecal samples from the feeding experiment were extracted and prepared with the  
570 same protocol as described above for the original GF and SPF mice. Targeted analysis method was  
571 used for detection of phenylalanine conjugates for both unlabeled and C<sup>13</sup> labeled molecules. The  
572 areas under the curves were extracted and used for ratio calculations.

573

574 **LC-MS Metabolomics Data Processing from PRISM and iHMP cohorts from the HMP2**  
575 **IBD Datasets.** The raw LC-MS data were acquired to the data acquisition computer interfaced to  
576 each LC-MS system and then stored on a robust and redundant file storage system (Isilon Systems)  
577 accessed via the internal network at the Broad Institute. Nontargeted data were processed using  
578 Progenesis Qlsoftware (v 2.0, Nonlinear Dynamics) to detect and de-isotope peaks, perform  
579 chromatographic retention time alignment, and integrate peak areas. Peaks of unknown ID were  
580 tracked by method, *m/z* and retention time. The novel conjugated bile acids were searched for by  
581 matching *m/z* in negative mode and subsequently verified using LC-MS/MS and synthetic standards  
582 of Phe-chol, Tyr-chol and Leu-chol from pooled samples (table S8).

583

584 **Statistical Analysis of HMP2 Metabolomics Data.** Prior to model fitting, raw metabolite  
585 abundances were median-normalized within sample and then log-transformed with a pseudocount of  
586 1. We used linear models implemented in R to associate metabolite abundances with IBD phenotype  
587 while controlling for clinical covariates. For the cross-sectional PRISM data, we treated categorical  
588 IBD diagnosis (UC, CD, and non-IBD control) as the phenotype of interest with “non-IBD” as a  
589 reference group. Age was included as a continuous covariate, while antibiotics, immunosuppressants,  
590 mesalamine, and steroids use were coded as binary covariates. The model was evaluated as follows  
591 using R's *lm* function:

592

593 
$$\text{metabolite} \sim (\text{intercept}) + \text{diagnosis}$$
  
594 
$$+ \text{age} + \text{antibiotic} + \text{immunosuppressant} + \text{mesalamine} + \text{steroids}.$$

595

596 The nominal *p*-values of the diagnosis coefficients for each metabolite were adjusted for multiple  
597 hypothesis testing using the Benjamini-Hochberg FDR method. A more sophisticated mixed-effects  
598 model was applied per-feature to the HMP2 metabolomics data to account for repeated measures  
599 over subjects and the multiple recruitment sites within the study. In addition, the transformed  
600 abundance of each metabolite was modeled as a function of a combined phenotype: diagnosis (as  
601 defined above) and dysbiosis state as a nested binary variable within each diagnosis (with non-  
602 dysbiotic as reference). The definition of “dysbiosis state” is presented in detail in the next section.

603 Model results were further adjusted for consent age as a continuous covariate and antibiotics use as a  
604 binary covariate. The mixed effects model was evaluated as follows using the *lme* function in R's *nlme*  
605 package [where (1 | subject) and (1 | recruitment site) indicate random effects for subject and  
606 recruitment site, respectively]:

607

608 | metabolite ~ (intercept) + diagnosis + diagnosis/dysbiosis\_+ antibiotic use + consent age + (1 |  
609 recruitment site) + (1 | subject)

610

611 Statistical significance (*p*-value) of metabolite-phenotype associations were assessed using Wald's  
612 test and corrected for multiple hypothesis testing as described above.

## 613 **Dysbiosis analyses**

### 614 Dysbiosis score

615 To identify samples with highly divergent (dysbiotic) metagenomic microbial compositions in the  
616 HMP2 dataset, a “dysbiosis score” was defined as in<sup>10</sup> based on Bray-Curtis dissimilarities to non-IBD  
617 metagenomes. First, a “reference set” of samples was constructed from non-IBD subjects by taking all  
618 samples after the 20th week after the subject's first stool sample. This was chosen since a subset of  
619 the non-IBD subjects at the start of their respective time series may not yet have overcome any  
620 gastrointestinal symptoms that triggered the initial visit to a doctor, though ultimately not caused by  
621 IBD. The dysbiosis score of a given sample was then defined as the median Bray-Curtis dissimilarity  
622 to this reference sample set, excluding samples that came from the same subject. To identify highly  
623 divergent samples, we then thresholded the dysbiosis score at the 90th percentile of this score for  
624 non-IBD samples. This therefore identifies samples with a feature configuration that has a <10%  
625 probability of occurring in a non-IBD subject. By this measure, 272 metagenomes were classified as  
626 dysbiotic. Samples from CD and UC subjects are overrepresented in the dysbiotic set, with 24.3% and  
627 11.6% of their samples classified as dysbiotic, respectively. A metabolite measurement was then  
628 defined as dysbiotic only if its paired metagenome is defined as dysbiotic according to the above  
629 definition. Only metabolomes with matched metagenomes were used in differential abundance testing  
630 (for UC, 12 dysbiotic and 110 non-dysbiotic metabolomes; for CD, 48 dysbiotic and 169 non-  
631 dysbiotic).

632

633 **Bile acid Gavage of Mice.** Eight-week-old Male C57BL/6J mice (Jackson Laboratory) were  
634 acclimated for 14 days and housed in groups of two throughout the duration of the experiment to  
635 mitigate cage effects. Mice were then dosed 2 times (24 hour) or 4 times (72 hour) (t = 0 hr, t = 24 hr,  
636 t = 48hr, t = 72hr) by oral gavage with either a mock control of corn oil without bile acids, or corn oil

infused Tyr-chol (500 mg/kg body weight), Leu-chol (500 mg/kg body weight), cholic acid (500 mg/kg body weight) or the control FXR agonist GW4064 (10 mg/kg body weight). Starting 3 hours after the last gavage (t = 75 hr, 72 hour treatment or t = 25, 24 hour treatment), mice were euthanized by CO<sub>2</sub> asphyxiation and samples were collected within a 6 hour period and snap frozen in a liquid nitrogen bath and stored at -80°C prior to analysis. All mice were handled in accordance with guidelines for the humane care and use of experimental animals, and the procedures used were approved by the University of California, San Diego Institutional Animal Care and Use Committee and the Salk Institute for Biological Studies Institutional Animal Care and Use Committee. Ileum and liver samples were used for qPCR.

646

**RT-qPCR Analysis of Downstream FXR Gene Expression.** Mouse liver and ileum segments were directly homogenized in TRIzol and total RNA isolated. cDNA was synthesized from 1µg of DNase-treated total RNA using Bio-Rad iScript Reverse Transcription supermix (#1708841) and mRNA levels of *Fgf15*, *Shp*, *Cyp7b1* and *Cyp7a1* were quantified by quantitative PCR with Advanced Universal SyBr Green Supermix (Bio-Rad, cat #725271). All samples were run in technical triplicates and relative mRNA levels were calculated by using the standard curve methodology and normalized to *36B4*. All primers are listed in the Supplementary Table S9.

654

## 655 References Methods

- 656 1. Tripathi, A. *et al.* Intermittent Hypoxia and Hypercapnia, a Hallmark of Obstructive Sleep  
657 Apnea, Alters the Gut Microbiome and Metabolome. *mSystems* **3**, e00020-18 (2018).
- 658 2. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research.  
659 *mSystems* **3**, e00031-18 (2018).
- 660 3. Cullen, T. W. *et al.* Antimicrobial peptide resistance mediates resilience of prominent gut  
661 commensals during inflammation. *Science* (80-. ). **347**, 170–175 (2015).
- 662 4. Integrative HMP (iHMP) Research Network Consortium, T. I. H. (iHMP) R. N. The Integrative  
663 Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods  
664 of human health and disease. *Cell Host Microbe* **16**, 276–89 (2014).
- 665 5. Amorim, P., Moraes, T., Silva, J. & Pedrini, H. InVesalius: An Interactive Rendering Framework  
666 for Health Care Support. in 45–54 (Springer, Cham, 2015). doi:10.1007/978-3-319-27857-5\_5
- 667 6. Casteleyn, C., Rekecki, A., Van der Aa, A., Simoens, P. & Van den Broeck, W. Surface area  
668 assessment of the murine intestinal tract as a prerequisite for oral dose translation from mouse  
669 to man. *Lab. Anim.* **44**, 176–83 (2010).
- 670 7. Shalapour, S. *et al.* Inflammation-induced IgA+ cells dismantle anti-liver cancer immunity.  
671 *Nature* **551**, 340–345 (2017).
- 672 8. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data.

- 673 *Nat. Methods* **7**, 335–6 (2010).
- 674 9. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq  
675 and MiSeq platforms. *ISME J.* **6**, 1621–4 (2012).
- 676 10. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel  
677 diseases. *Nature* **569**, 655–662 (2019).
- 678 11. Protsyuk, I. *et al.* 3D molecular cartography using LC–MS facilitated by Optimus and 'ili  
679 software. *Nat. Protoc.* **13**, 134–154 (2017).
- 680 12. Kenar, E. *et al.* Automated Label-free Quantification of Metabolites from Liquid  
681 Chromatography–Mass Spectrometry Data. *Mol. Cell. Proteomics* **13**, 348–359 (2014).
- 682 13. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for  
683 processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC*  
684 *Bioinformatics* **11**, 395 (2010).
- 685 14. Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis Chemical  
686 Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**,  
687 211–221 (2007).
- 688 15. Scheubert, K. *et al.* Significance estimation for large scale metabolomics annotations by  
689 spectral matching. *Nat. Commun.* **8**, 1494 (2017).
- 690 16. Hartmann, A. C. *et al.* Meta-mass shift chemical profiling of metabolomes from coral reefs.  
691 *Proc. Natl. Acad. Sci. U. S. A.* **114**, (2017).
- 692 17. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences  
693 per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl**, 4516–22 (2011).
- 694 18. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns.  
695 *mSystems* **2**, (2017).
- 696 19. Lozupone, C. & Knight, R. UniFrac : a New Phylogenetic Method for Comparing Microbial  
697 Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities.  
698 *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
- 699 20. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and  
700 workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–72 (2006).
- 701 21. Mirarab, S., Nguyen, N. & Warnow, T. SEPP: SATé-enabled phylogenetic placement. *Pac.*  
702 *Symp. Biocomput.* 247–58 (2012).
- 703 22. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial  
704 communities. *Appl. Environ. Microbiol.* **71**, 8228–35 (2005).
- 705 23. Ezawa, T., Jung, S., Kawashima, Y., Noguchi, T. & Imai, N. Ecological Base-Conditioned  
706 Preparation of Dipeptides Using Unprotected  $\alpha$ -Amino Acids Containing Hydrophilic Side  
707 Chains. *Bull. Chem. Soc. Jpn.* **90**, 689–696 (2017).
- 708 24. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural

Products Social Molecular Networking. *Nat. Biotechnol.* **34**, (2016).

25. Morton, J. T. *et al.* Learning representations of microbe–metabolite interactions. *Nat. Methods* 1–9 (2019). doi:10.1038/s41592-019-0616-3
26. McDonald, J. A. K. *et al.* Evaluation of microbial community reproducibility, stability and composition in a human distal gut chemostat model. *J. Microbiol. Methods* **95**, 167–174 (2013).

### Supplementary Data

#### **Overall Microbiome and Metabolome Relationships.**

A broad overview of data relationships was first assessed through principal coordinates analysis (PCoA) using the Bray-Curtis dissimilarity matrix (metabolome) and UniFrac distance (microbiome) (Extended Data Fig. 1a). The metabolome data was most strongly influenced by organ source (Extended Data Fig. 1b,c). When plotted by organ, four distinct metabolome clusters emerged: the gastrointestinal (GI) tract, epidermal swabs, blood rich organs (lung, heart, spleen, and blood itself), and a cluster of all other visceral organs (Extended Data Fig. 1a,b). We further collected 16S inventories to understand the spatial pattern of bacterial colonization in the mice. As expected, the microbiome data was dictated by colonization status. GF mice and sterile organs in SPF mice clustered tightly with background sequence reads from blanks (reflecting their sterility), whereas colonized organs within the SPF mice clustered apart from these samples (Extended Data Fig. 1a,b). Notable separation of certain organ systems was observed in the microbiome of SPF mice, including a distinct grouping of the GI tract (including the esophagus) and clustering of the vagina and cervix samples (Extended Data Fig. 1a,b). To quantify the effect of microbial colonization on the metabolomic data, the Bray-Curtis dissimilarity was calculated between the MS<sup>1</sup> data of GF and SPF mice, then compared to the within group variation for all paired sample locations with statistical significance being determined by Mann-Whitney U-test. The strongest separation between the metabolomic data was present in stool, followed by the cecum, other regions of the GI tract, and samples from the surface of the animals including ears and feet (Extended Data Fig. 1c). Thus, the major molecular signatures distinguishing colonized and GF mice were present in the gut and epidermis with particularly strong effects in the stool, cecum and ileum. The liver also had signatures suggestive of metabolomic differences between the GF and SPF mice, but this was not significant compared to the within individual variation (Extended Data Fig. 1c).

The 16S rRNA gene microbiome profiles of the GI tract were dominated by Bacteroidales clade S24-7, Firmicutes, *Lactobacillus* and *Akkermansia muciniphila* (Extended Data Fig. 1d). Large changes in microbial profiles were observed traversing the GI tract. The esophagus, stomach and duodenum had relatively similar profiles, but a dramatic shift in the jejunum with the expansion of *Lactobacillus* and *A. muciniphila* and a decrease in the relative abundance of Bacteroidales S24-7 was evident. The community transitioned again through the ileum with a further expansion of

744 *Lactobacillus*. At the cecum an abrupt transition was observed with a reduction of *Lactobacillus* and  
745 increase in the relative abundance of Firmicutes (Extended Data Fig. 1d), this community was largely  
746 maintained through the colon until the stool, where the Firmicutes were reduced (Extended Data Fig.  
747 1d).

748 **Unique molecules from the microbiome.** Molecular networking paired with statistical analysis  
749 enabled identification of molecules unique or enriched between the two groups of mice. These  
750 included bile acids, flavonoids, triterpenoid saponins, and urobilins (Extended Data Fig. 1-4). The  
751 soyasaponins and flavonoids were prevalent, diverse and differentially abundant between the two  
752 groups of mice. These compounds were sourced from the mouse chow that had a dominant soybean  
753 component. A cluster of 76 connected nodes in the molecular network representing soyasaponins  
754 was found in both GF and SPF mice and their food pellets, but these clusters were enriched in nodes  
755 from the GI tract of GF mice (Extended Data Fig. 2). This molecular family contained a variety of  
756 unique soyasaponins all comprised of the core soyasapogenol triterpenoid backbone, but with  
757 different glycosylations and hydroxylations. Soyasaponins were present throughout the GI tract of GF  
758 mice, including the stool sample, but in SPF mice they disappeared upon passage into the cecum  
759 (Extended Data Fig. 2). Conversely, there was a separate cluster only found in SPF mice that was  
760 annotated as soyasapogenols, which represent the triterpenoid backbone of soyasaponin without  
761 glycosylation (Extended Data Fig. 2). 3D-molecular cartography showed that soyasaponin I was  
762 abundant throughout the GI tract of GF mice, particularly the cecum, colon and stool, but was absent  
763 from these organs in SPF animals. In direct contrast, soyasapogenol was not found at all in GF  
764 animals, but was detected in the cecum of the SPF mice through to the stool. This differing presence  
765 of the glycone and aglycone forms indicates that cecal microbial activity was responsible for the  
766 metabolism of soyasaponin into soyasapogenol by removal of the saccharides (Extended Data Fig.  
767 2). The abundance of soyasapogenol E ( $m/z$  457.36) was then regressed against the microbiome  
768 data for significant associations between this metabolite and microbial operational taxonomic units  
769 (OTUs) (Bonferonni corrected p-value for 195 OTUs  $p < 2.6 \times 10^{-4}$ ). The Firmicute *Allobaculum* sp.  
770 (Pearson's  $r = 0.491$ ) was significantly correlated to the abundance of soyasapogenol E; the only  
771 cultured representative of this genus contains the  $\beta$ -glucosidase enzyme known to perform  
772 deglycosylation of plant natural products.

773 Microbiome breakdown of plant flavonoids was also observed (Extended Data Fig. 3). In the  
774 mouse chow, glucuronides and aglycone flavones and isoflavones were detected, but not their  
775 sulfated forms. Because many isomeric forms of flavonoids exist that cannot be differentiated with our  
776 MS/MS methods, we focused on molecular changes in the predominant soybean isoflavonoids  
777 daidzein, genistein and glycitein, because they have characteristic MS/MS signatures. In the GF mice,  
778 3D-molecular cartography showed that the glucuronidated and sulfated isoflavonoids were detected

779 throughout the GI tract from the stomach through to the stool, indicating they pass through the GI tract  
780 intact. In SPF mice, however, these same glucuronides and sulfides were undetectable in the distal GI  
781 tract. The aglycones were present in both the GF and SPF mice, but more abundant in the distal GI  
782 tract of GF animals (Extended Data Fig. 3, Mann-Whitney U-test,  $p < 0.05$ ). Because the aglycones  
783 were detected in both groups, host and microbial enzymes (or chemical processes) could have been  
784 responsible for the deglycosylation; however, the complete removal of the sugars and sulfates in the  
785 SPF mice indicated that the microbiota significantly enhanced this process. Furthermore, in the cecum  
786 of the SPF mice, the aglycone isoflavonoids were depleted and in some cases no longer detectable  
787 through to the stool samples, indicating that further metabolism of these compounds was occurring in  
788 the cecum and colon due to the presence of bacteria.

789 The production of secondary bile acids was also prevalent in SPF mice, but not GF mice.  
790 Deoxy- and keto-forms (dehydrogenated) of cholic acid were abundant in the distal GI tract of SPF  
791 mice but absent from GF mice (Extended Data Fig. 4). In contrast, the primary bile acid  
792 tauromuricholic acid was abundant throughout the gut of GF mice but was depleted in the distal GI  
793 tract of SPF mice. Muricholic acid was also exclusive to the guts of SPF mice but was found in the  
794 liver of sterile animals (Extended Data Fig. 4).

795

796 **MS/MS Annotation of Novel Conjugated Bile Acids.** Analysis of the unique nodes in SPF mice  
797 related to glycocholic acid (Fig. 2a) led to the discovery of the unique conjugation with different amino  
798 acids. The major core fragment of cholic acid in all conjugated bile acids is shown in Extended Data  
799 Figure 5d at mass  $m/z$  337.25. This represents the core steroid backbone of cholic acid with loss of  
800 the amino acid conjugate and all hydroxyl groups. In the new Tyr, Phe and Leu conjugates the  
801 difference in mass of a whole amino acid can be seen from the parent ion and this fragment.  
802 Furthermore, this amino acid ion appears in the lower  $m/z$  range of the spectrum as the whole amino  
803 acid plus a hydrogen ion ( $H^+$ ). Further verification of these molecules comes from the presence of  
804 unique immonium ions, a characteristic of peptide fragmentation, which are seen in the lower mass  
805 range corresponding to each of the three amino acids (Extended Data Figure 5d, table S2).

806

807 **Amino Acid Conjugate Synthesis and Validation.** Both cholic and muricholic forms of the three  
808 novel amino acid conjugates and an isoleucine conjugate were chemically synthesized and verified  
809 using nuclear magnetic resonance spectroscopy (NMR spectra below). Polarity and MS/MS  
810 fragmentation patterns of these compounds were subsequently analyzed and the higher hydrophilicity  
811 of muricholic acid forms were validated by earlier retention times for all four synthesized compounds  
812 (Extended data Fig. 5b,c). MS/MS patterns of muricholic and cholic acid forms were identical and the  
813 spectra from the SPF mouse duodenum were subsequently verified to match these synthetic  
814 compounds by molecular networking, retention time analysis and MS/MS matching (Extended data



Fig. 5). In the mouse jejunum sample the extracted ion chromatogram for leucocholic acid (m/z 522.3700) contained a single peak that most closely matched leucocholic acid, however, there was a small shoulder on this peak indicating that it cannot be ruled out that some isoleucocholic acid may be present (Extended data Fig. 5c). MS/MS patterns of synthetic standards and novel bile acids from mouse gut samples showed high similarity (Extended data Fig. 6a)

**Bile Acids in Murine Portal and Peripheral Blood.** An additional 4 SPF and 6 GF female mice of the same strain analyzed for the initial study on the microbial metabolome were raised for analysis of blood. Portal blood and peripheral blood were sampled as described in the methods section and analyzed with the same LC-MS/MS protocols as the original animals. Parent masses for the Phe, Tyr and Leu conjugated microbial bile acids that were searched for in the GNPS molecular network were not found (Extended data Fig. 6c). The conjugated bile acid molecular family was further inspected for the presence of these compounds but was also negative for the presence of the novel conjugates in either peripheral or portal blood samples from either mouse group. The host conjugated taurocholic acid and glycocholic acid were however, found in both blood types of both murine groups.

**Synthesis of Novel Conjugate Bile Acids by *Clostridium bolteae*.** After finding a strong association between all three novel conjugates and a *Clostridium* sp. in mice fed high fat diet<sup>25</sup> (Extended data Fig. 7, table S3), twenty isolates of human gut bacteria were cultured in fecal culture media and screened for the production of these compounds using the same extraction and LC-MS/MS methods described for the mouse organ analyses. Using GNPS integrated with mzMine feature finding, Phe-chol was detected in the extracts from three separate *Clostridium* strains, but at very low intensity. Only *C. bolteae* had produced the molecule clearly at a level at least 3x the abundance of the background extracted ion chromatogram trace. Thus, using the more sensitive targeted and quantitative assay we subsequently repeated these experiments with two isolates of *C. bolteae* and validated the production of both Phe-chol and Tyr-chol in the culture extracts (Extended data Fig. 8). More of the tyrosine conjugate was made than (~20 ng/ml) the phenylalanine conjugate (~7 ng/ml). Further validation was provided using media supplemented with <sup>13</sup>C labeled phenylalanine added to the media. This labeled amino acid was incorporated into the Phe-chol produced by *C. bolteae* WAL-14578 demonstrating that free amino acids from the media can be used for the conjugation and providing direct evidence that these bile acids are made by microbes (Fig. 3, Extended data Fig. 8).

**Detection of Novel Bile Acid Conjugates in HMP2 dataset.** Phe-chol, Leu-chol and Tyr-chol were detected in the HMP2 dataset with negative ion mode (table S8). The statistical testing for differences between inflammatory bowel disease patients in HMP2 are as follows: IBD patients (Fig. 3c, PRISM

dataset, FDR-corrected p-value (q-value) from Wald's test of linear effects model of Leu = 0.03, Tyr = 0.0074 and Phe = 0.004, control non-IBD n = 34, CD n = 68, and UC n = 53). Furthermore, they were enriched in CD dysbiosis (HMP2 dataset q-value, Phe = 0.0003, Tyr = 0.007, Leu =  $9.0 \times 10^{-5}$ , n=48 CD-dysbiotic, n= 169 CD non-dysbiotic) but not statistically different in UC dysbiosis (q=1.0, 0.8, 0.9 for Phe, Tyr, Leu-cholate amides, n=12 UC dysbiotic, n=110 UC -non-dysbiotic) and not in non-IBD (q=0.4, 0.5, 0.5 for Phe, Tyr, Leu-cholate amides, n=15 non-IBD-dysbiotic, n=107 non-IBD-non-dysbiotic, Wald's test).

858

**Sequencing of Fecal Cultures Exposed to Novel Bile Acid Conjugates.** In the batch culture experiment where an actively growing fecal culture was exposed to the novel conjugated bile acids and other control molecules, the microbiome of the culture media was sequenced using 16S rRNA amplicon sequencing after 24 hours. The data was processed with the Qiita pipeline and the resultant cultures were analyzed for changes in the microbiome structure due to conjugated bile acid exposure. There was no change in the microbiome alpha-diversity when cultured in the presence of any bile acids added to the media compared to the mock control. The Shannon diversity of the community decreased over time, but this was not different than the mock control with no bile acids added (Extended data Fig. 8c).

868

**Quantification of Bile Acids.** The concentration of the new bile acids in the mouse gut samples was quantified in negative-mode using the targeted method by comparison to the standard curves measured of each molecule in the various tissue samples spiked into the GF mice samples. The calculation was then normalized to the initial g/tissue collected (masses of samples in table S1) and the dilution through extraction and mass spectrometry analysis (Table S4, Extended data Fig. 8).

874

**Matrix Effects on Novel Conjugated Bile Acids.** Standards of the novel conjugated bile acids were added to the gut and other samples of germ-free mice to determine the matrix effects on each compound in the targeted method using a triple-quad mass spectrometer (see methods). Although some ion suppression (64% for the phenylalanine conjugated cholic acid in the duodenum) and ion enhancement (135% for the leucine conjugated cholic acid in the duodenum) were observed, the average matrix effects using the positive mode method was 100% (table S6). Calculated matrix effect values were in the range of 80 to 120%, indicating low matrix effects in the ESI positive ion source on these bile acid compounds. Matrix effect was stronger using the negative-mode targeted method, particularly in the blood samples (table S7) but the limit of detection was 11x lower than positive mode thus it was used for quantification with matrix matched calibration.

885

886 **RT-qPCR analysis of downstream FXR effector genes.** The gene expression of *Fgf15* and *Shp* in  
 887 the ileum and *Shp*, *Cyp8b1* and *Cyp7a1* in the liver of mice gavaged with bile acids of interest were  
 888 analyzed using quantitative reverse transcriptase-PCR analysis. The expression levels were  
 889 normalized to the cellular housekeeping gene ribosomal phosphoprotein PO (36B4). Mice were  
 890 sacrificed at both 24 hr (Extended data Fig. 9) and 72 hr post (Fig 3.e) gavage. At the 24-hr time point  
 891 expression of the downstream FXR effectors *Fgf15* were both significantly elevated ( $p<0.05$ ) after  
 892 gavage with Tyr-chol in the ileum, significance was also reached for *Shp* with cholic acid (CA) and the  
 893 GW4064 synthetic agonist. In the liver at 24hrs, *Cyp8b1* and *Cyp7a1* were significantly reduced in  
 894 expression in the Tyr-chol, Leu-chol and cholic acid treatments (Extended data Fig. 9). *Shp* signaling  
 895 was not significantly affected at this time point. At 72 hrs post gavage, ileum *Fgf15* and *Shp* signaling  
 896 were significantly increased for the Tyr-chol, Leu-chol, and CA groups (Fig. 3e). Liver expression of  
 897 *Shp* was also significantly elevated, but only in the Tyr and Leu conjugates. The bile acid synthesis  
 898 enzymes *Cyp8b1* and *Cyp7a1* were both significantly reduced compared to the corn oil control in Tyr-  
 899 chol, Leu-chol and CA gavages (Fig. 3e).

900

# 901 **Supplementary Tables**

902 Table S1. Sample information from GF and SPF mice analyzed in this study.

903

904 Table S2. Mass spectrometry details and ions of interest for identification of novel conjugated bile  
 905 acids.

Compound	Exact Mass	Observed Mass	Charge	Immonium Ion	Amino acid fragment	Other diagnostic fragments
Phenylalanocholeic acid	555.3559	556.362	H+	120.0816	166.0862	337.2525, 319.2420, 227.1398
Tyroscholeic acid	571.3509	572.356	H+	136.0758	182.081	337.2525, 319.2420, 227.1398
Leucocholeic acid	521.3716	522.379	H+	86.0977	132.1002	337.2525, 319.2420, 227.1398

906

907 Table S3. Deblurred read, taxonomic assignment and Pearson's r correlation for bacterial assigned  
 908 sequence variants (ASVs) with the three novel conjugated bile acids in the HFD feeding experiment  
 909 (n=12).

910

911 Table S4. a) Quantification of novel conjugated bile acids in mouse SPF gut samples (n=4) and  
 912 standard deviation of the mean across the different organ samples. b) Number of samples that had  
 913 values above LOD included in the calculations for Table S4a. c) Limit of detection of novel conjugated  
 914 bile acids with different background matrices.

915 a)

Organ	Mean nmol/g tissue		Standard Deviation	
	Tyr	Phe	Tyr	Phe
Jejunum	114.09	147	79.01	99.91
Ileum	56.03	83.56	57.85	81.33
Cecum	<LOD	4.74	0	3.38
Colon	<LOD	11.61	0	12.21

916

917 b)

	Tyr	Leu	Phe
Cecum (n=24)	0	0	11
Colon (n=24)	0	0	9
Ileum (n=24)	18	16	18
Jejunum (n=24)	18	18	18

918

919 c)

Organ	Limit of Detection (ppb)		
	Tyr	Ile	Phe
Jejunum	2.70	3.12	1.74
Ileum	2.73	1.89	3.01
Cecum	3.30	2.45	1.78
Colon	7.25	5.00	1.19

920

921 Table S5. MRM transitions and mass spectrometry details for targeted method. a) Negative mode b)  
922 positive mode. Other details of the mass spectrometry method are also provided

923 a)

#### Negative Mode

	Q1	Q3	CE	Tube lens	tr (min)
Leu	520.4	130.2	44	100	6.31
Leu	520.4	458.2	38	100	6.31
Phe	554.4	147.2	42	100	6.38
Phe	554.4	164.1	42	100	6.38
PheC13	555.4	148.2	44	100	6.38
PheC13	555.4	165	42	100	6.38
Tyr	570.4	119.1	48	100	5.96
Tyr	570.4	179.9	38	100	5.96
Tyr10	580.4	190	41	150	5.96
Tyr10	580.4	535.2	38	150	5.96

Spray voltage 2500  
 Vaporizer Temperature 267  
 Sheath gas pressure 39  
 Aux gas pressure 33  
 Capillary temperature 355

924

925 b)

#### Positive Mode

	Q1	Q3	CE	Tube lens	
Leu	522.4	337	25	170	9.36
Leu	522.4	468.1	19	170	9.36
Phe	556.4	337.1	23	190	9.6
Phe	556.4	389	37	190	9.6
Tyr	572.4	337.1	20	160	8.5
Tyr	572.4	518	17	160	8.5

Spray voltage 3000  
 Vaporizer Temperature 350  
 Sheath gas pressure 39  
 Aux gas pressure 33  
 Capillary temperature 380

926

927 Table S6. Matrix effect values for different sample types in positive ionization mode for the conjugated  
 928 bile acids. The effects are expressed as a percentage from the analyzed chemical standard.

	Tyr-Chol	Leu-Chol	Phe-Chol
Stool	124	92	95
Jejunum	99	135	69
Ileum	83	130	87
Duodenum	124	128	85
Cecum	96	91	64
Colon	123	113	95
Stomach	83	96	80

929

930 Table S7. Matrix effect values for different sample types in negative ionization mode using the  
 931 targeted method for the conjugated bile acids. The effects are expressed as a percentage from the  
 932 analyzed chemical standard.

	Tyr-Chol	Leu-Chol	Phe-Chol
Ileum	27	52	48
Cecum	66	79	77
Colon	21	23	22

Jejunum	32	57	55
Fecal	67	83	86
Blood	4	22	25

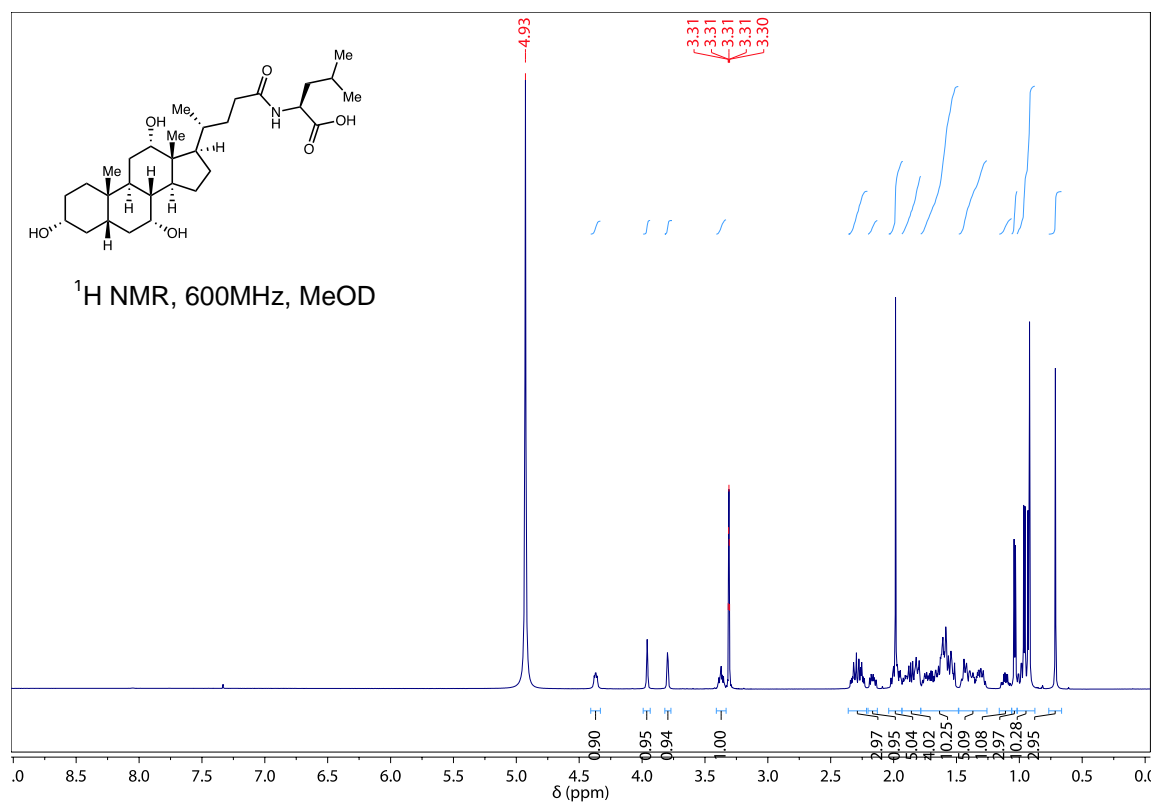
Table S8 - Mass spectrometry and retention time characteristics of the Phe, Tyr and Leu conjugated bile acids.

	Compound	Exact Mass	Observed Mass	Retention Time	Charge	Immonium Ion	Amino acid fragment	Other diagnostic fragments
Pos mode	phenylalanocholeic acid	555.3559	556.362	5.9 min	H+	120.0816	166.0862	337.2525, 319.2420, 227.1398
	tyrosocholeic acid	571.3509	572.356	5.3 min	H+	136.0758	182.081	337.2525, 319.2420, 227.1398
	leucocholeic acid	521.3716	522.379	5.8 min	H+	86.0977	132.1002	337.2525, 319.2420, 227.1398
Neg mode	phenylalanocholeic acid	555.3559	554.3491	5.9 min	H-	NA	164.0709	302.2722, 221.2677
	tyrosocholeic acid	571.3509	570.3499	5.3 min	H-	NA	180.066	302.2722, 220.9721
	leucocholeic acid	521.3716	520.3646	5.8 min	H-	NA	130.0864	302.2722, 221.0867

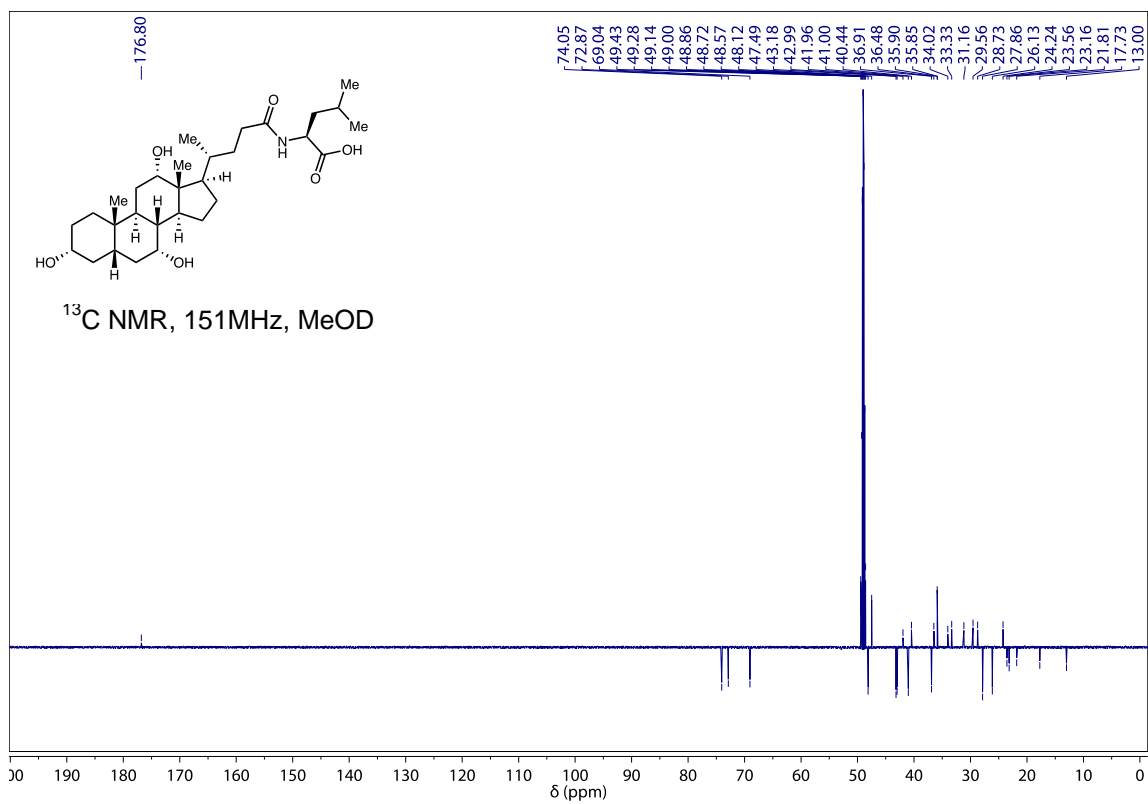
Table S9 – Primers used for qPCR quantification of Fgf15, Shp, Cyp8b-1 and Cyp7a-1 compared to the 36b4 housekeeping gene control.

Primer	Sequence
m36b4-F	ACCTCCTTCTTCTTCCAGGCTTT
m36b4-R	CCCACCTTGTCTCCAGTCTTT
mFgf15-F	GCCATCAAGGACGTCAGCA
mFgf15-R	CTTCCTCCGAGTAGCGAATCA
mShp-F	CTACCCTCAAGAACATTCCAGG
mShp-R	CACCAGACTCCATTCCACG
mCyp8b1-F	GAAGTCAACCAGGCCATGCT
mCyp8b1-R	AGGAGCTGGCACCTAGACT
mCyp7a1-F	CATCTCAAGCAAACACCATTCC
mCyp7a1-R	TCACTTCTTCAGAGGCTGGTTTC

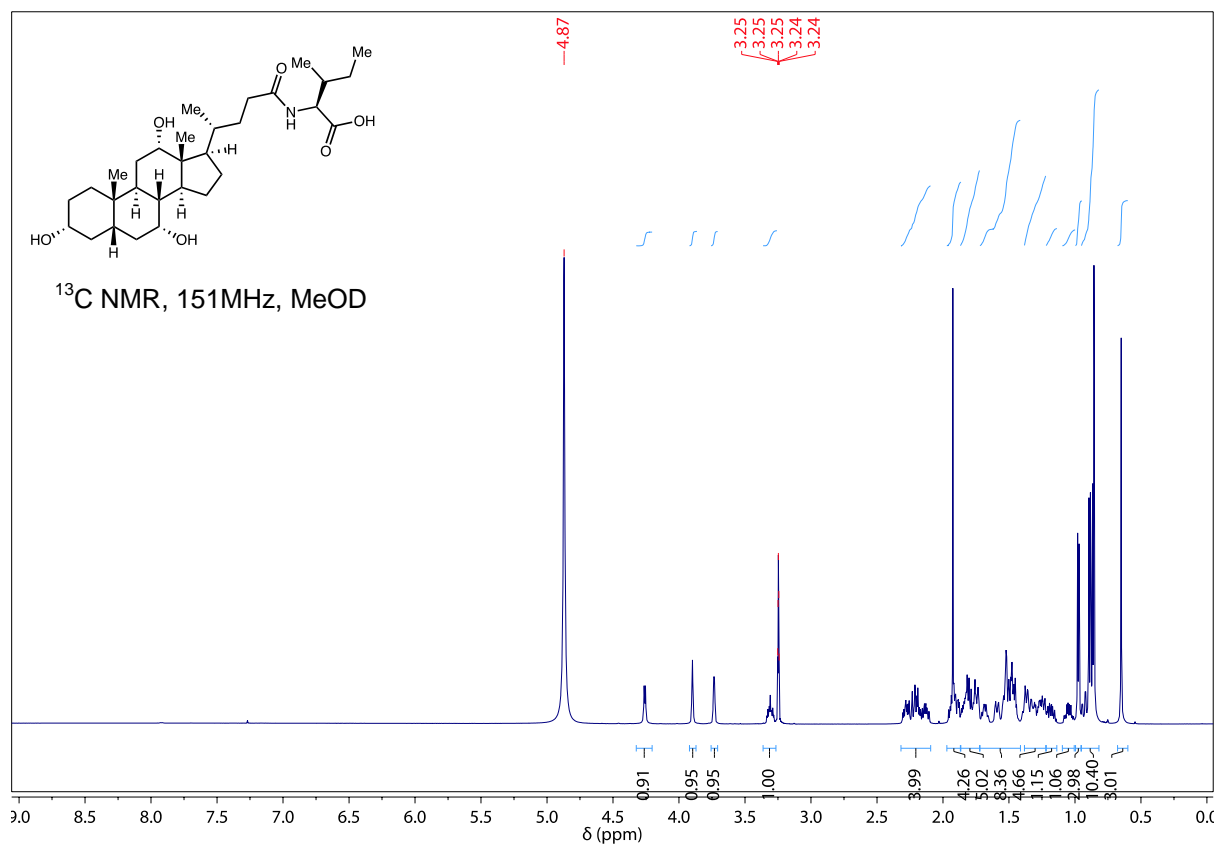
941 **Supplemental Data: NMR Spectra.** NMR spectral characterization of synthesized Leu-chol,  
942 Isoleu-chol, Phe-chol and Tyr-chol.



943

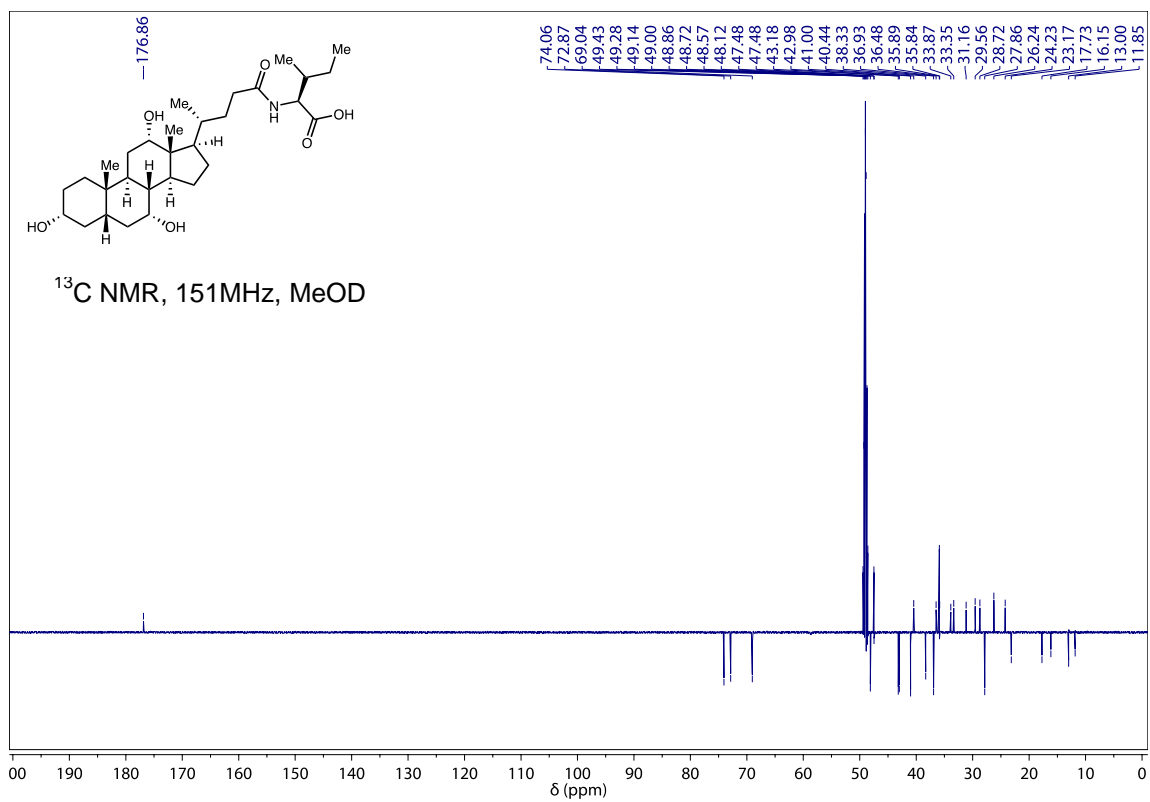


944

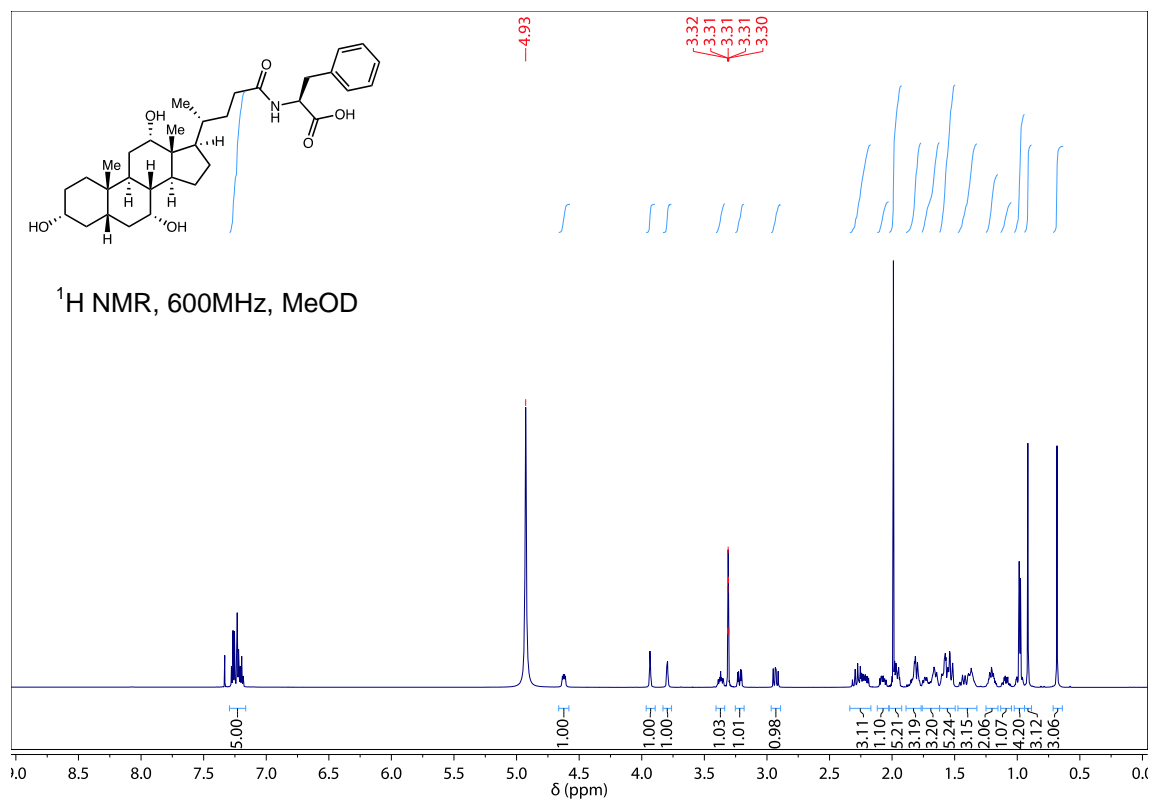


945

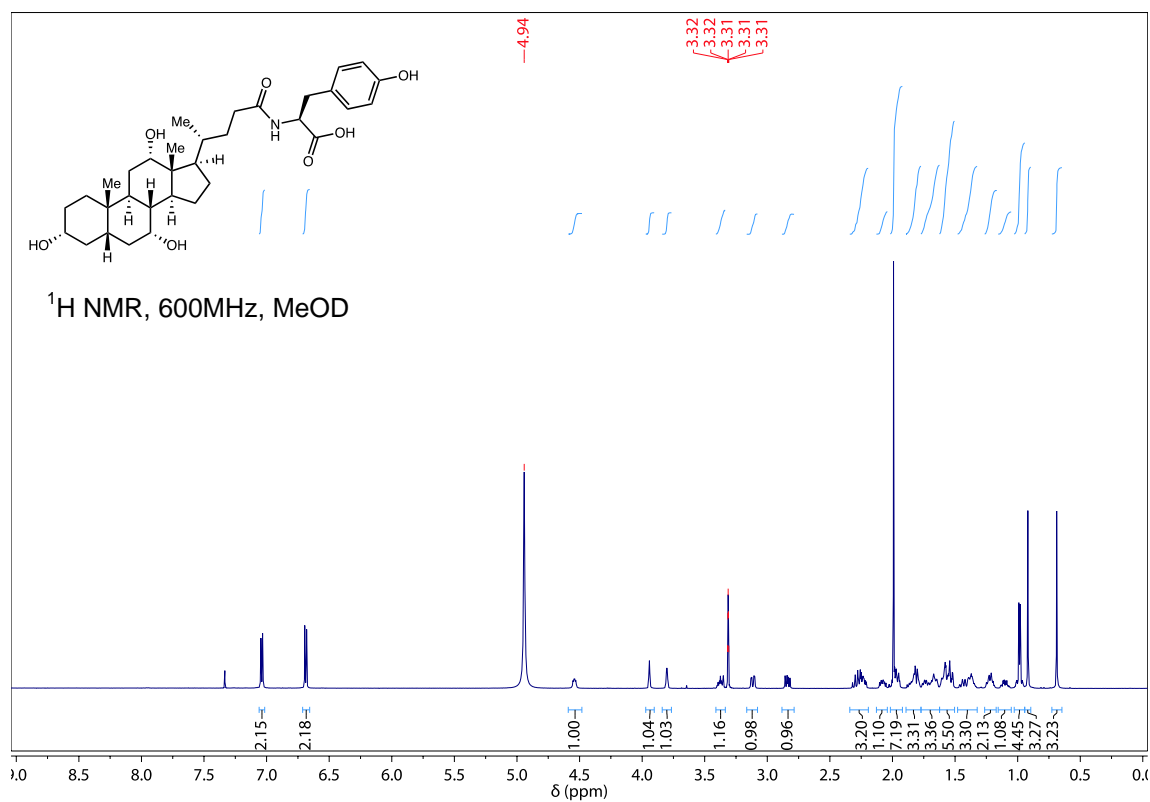
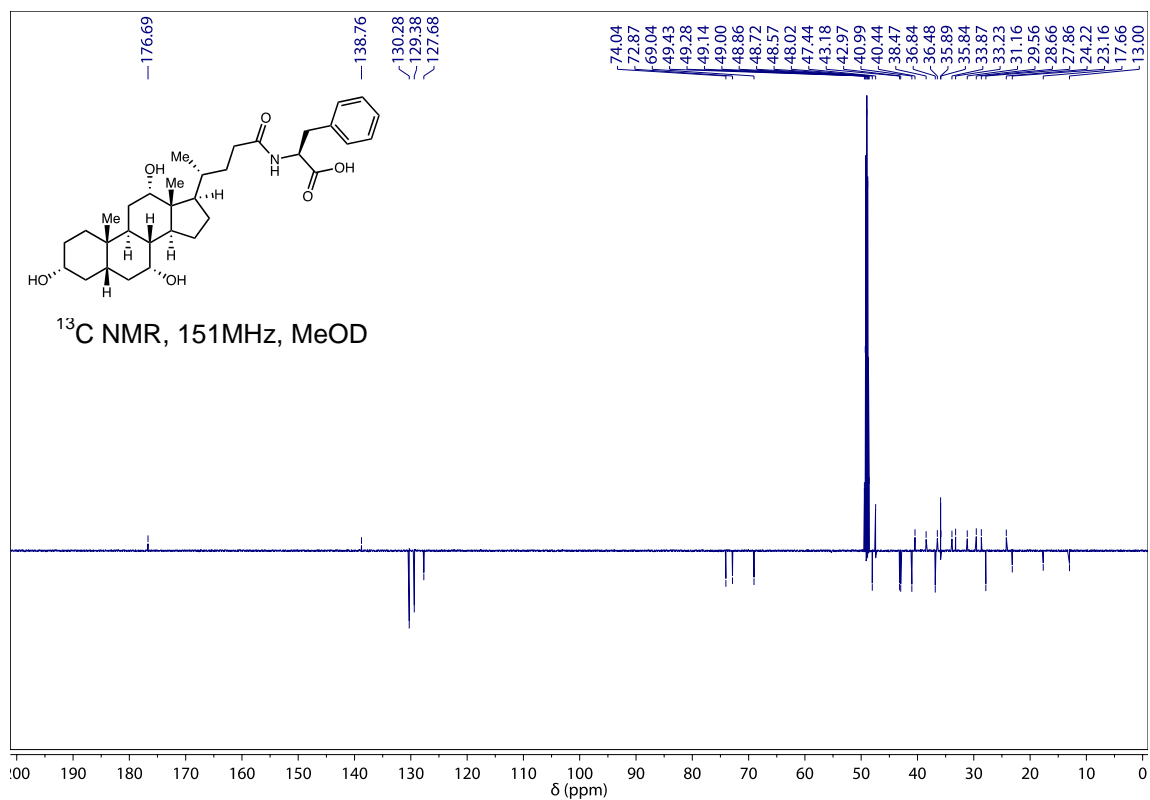


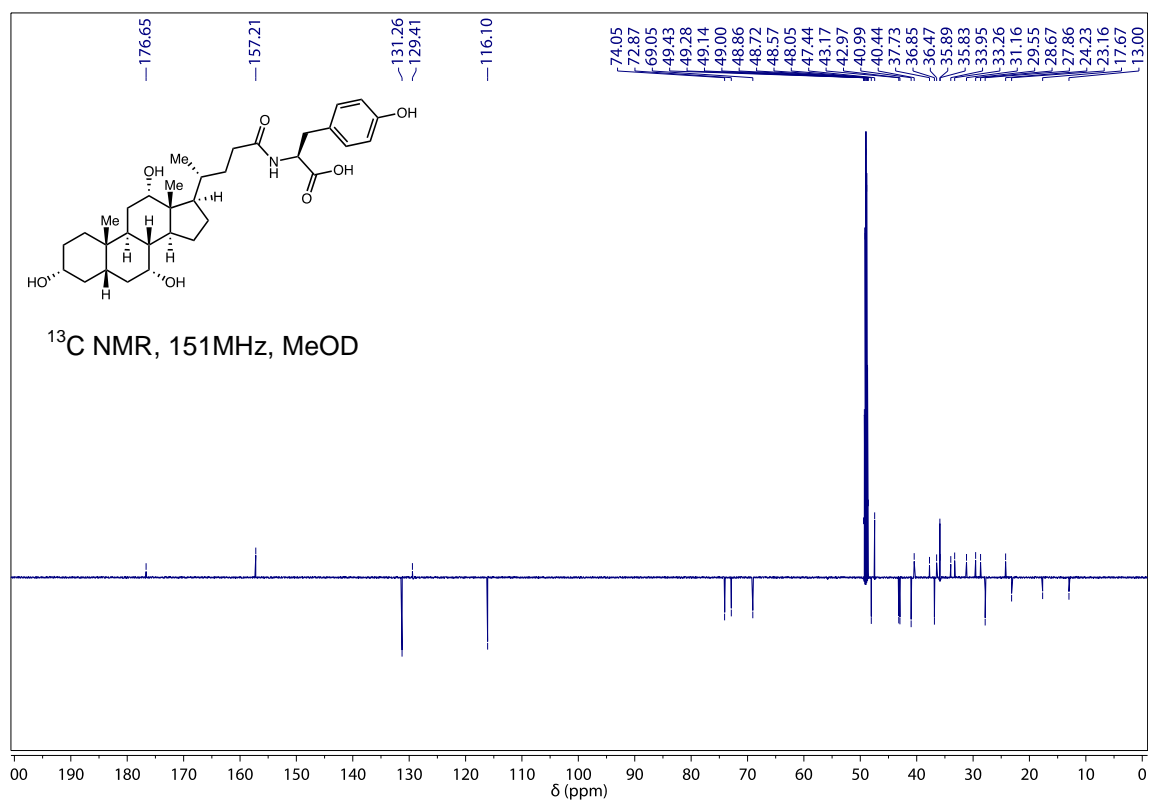


946



947





951  
952

### **Supplementary 3D Mouse Model**

Provided in the SI are .stl files that comprise the 3D mouse model for 3D-molecular cartography mapping. These include a full mouse model, the liver only, the GI tract only, and the GI tract without the liver. Also included are x,y,z coordinates that will enable mapping of multi-omics data to locations of interest on all four .stl files.